U. S. NAVAL PERSONNEL RESEARCH ACTIVITY

SAN DIEGO, CALIFORNIA 92152

TECHNICAL BULLETIN STB 67-18 FEBRUARY 1967

AD 659961

THE APPLICATION OF DECISION THEORY AND SCALING METHODS TO SELECTION TEST EVALUATION

Ervin W. Curtis

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

NPRA

SAN DIEGO CALIFORNIA

AN ACTIVITY OF THE BUREAU OF NAVAL PERSONNEL

THE APPLICATION OF DECISION THEORY AND SCALING
METHODS TO SELECTION TEST EVALUATION

Ervin W. Curtis

February 1967

IER16000001
Technical Bulletin STB 67-18

Submitted by

R. Rimland, Ph.D., Director, Personnel Measurement Research Department

Approved by

E. I. Jones, Ph.D., Technical Director, Acting
G. W. Watson, Commander, USN
Officer in Charge

Distribution of this document is unlimited

U. S. Naval Personnel Research Activity
San Diego, California 92152

ACKNOWLEDGEMENT

This research is based on a doctoral dissertation completed at the University of Oregon in 1965 under Dr. Leona Tyler, Chairman of the Thesis Committee. The study was done at the U. S. Naval Personnel Research Activity, where the author is a research psychologist.

The author wishes to express his gratitude to Dr. Tyler for her help and guidance, and to the U. S. Naval Personnel Research Activity for its support and cooperation during this endeavor.

BRIEF

The problem investigated in this study is that of evaluating psychological tests as aids to the selection of personnel for training and jobs. When an institution uses a test for the purpose of personnel selection, some estimate of its value as a decision-making tool is needed by psychologists and management. The conventional approach to test evaluation, namely, correlational analysis, ignores three important situational factors: how well the institution could do by chance (commonly called the "base rate"), the proportion to be selected from the population (the selection ratio), and the institutional gains and losses resulting from correct decisions and incorrect decisions.

A method based on statistic decision theory was developed which handles these factors explicitly and systematically. The method, as presented, is restricted to the dichotomous (or dichotomized) criterion case and does not rely on the correlation coefficient as an index of association between the test and the criterion. The decision-theoretic method involves the construction of a payoff matrix corresponding to the contingency table relating the test to the criterion. The cell frequencies are weighted in a utility equation by the payoff values (utilities) in the corresponding cells of the payoff matrix. This utility equation represen a new test evaluation index that directly expresses the utility of the te to the institution using it.

Also presented is a method based on Brogden's publications on thi problem. It involves the comparison of criterion groups, e.g., satisfactory and unsatisfactory, in terms of their utility to the institution using the selection test. It is called the "utility function" method since the criterion is converted to a utility scale.

The three methods (correlational, decision-theoretic, and utility function) were compared with tests used to select students for technical schools in the U. S. Navy. Scaling techniques were developed for the measurement of values inherent in the Navy situation. Specifically, the graduate-fail criterion was translated to a utility scale and the corresponding job areas were scaled on need (or the relative utility of graduates to the Navy). Using scale values obtained for the job areas, a payoff matrix was constructed for each school on the assumption that the currently used test cutoffs are optimal.

The three methods led to quite different indications regarding the utility of the selection tests evaluated. The decision-theoretic and utility function methods agreed in terms of the proportion improvement over chance prediction provided by the tests, while the correlational method tended to underestimate this proportion. In terms of utility, the decision-theoretic method indicated the tests were worth much more to the Navy than did the other two methods.

In addition to the above, the following conclusions were stated:

(1) Statistical decision theory is well suited to the usual selection testing situation. (2) Psychological scaling methods provide a solution for the measurement of values required in the application of the decision-theoretic approach to test evaluation. (3) Supplementation of correlational analysis of tests with decision-theoretic analysis is likely to lead to new insights into the utility and use of tests for personnel decisions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

STATEMENT OF THE PROBLEM AND THEORETICAL BACKGROUND

The problem investigated in this study is that of evaluating psychological tests as aids to the selection of personnel for training and jobs. When an institution uses a test for the purpose of personnel selection, some estimate of its value as a decision-making tool is needed by psychologists and management. The conventional approach to test evaluation, namely, correlational analysis, ignores three important situational factors: how well the institution could do by chance (commonly called the "base rate"), the proportion to be selected from the population (the selection ratio), and the institutional gains and losses resulting from correct decisions and incorrect decisions.

In an attempt to contribute to more adequate test evaluation, three tasks are undertaken in this study:

(1) Demonstration of the need for a new approach to selection test evaluation.

(2) Development of a mathematically rigorous yet practical approach to selection test evaluation which explicitly utilizes information about the base rate, selection ratio, and institutional gains and losses.

(3) Empirical tryout of the test evaluation approach developed in this study.

It is assumed throughout that to "evaluate" a test means to determine its value for a specific decision in a specific applied situation.

Background of two Diverse Approaches to Test Evaluation

The Conventional Approach

Personnel tests are typically evaluated by determining the correlation between the test and a criterion, usually some measure of performance. The resultant coefficient is commonly called the validity coefficient. Several indices have been developed for interpreting validity coefficients; the one having the longest history is the "index of forecasting efficiency," E:

$$E = 1 - \sqrt{1 - r^2},$$

where r is the correlation between the predictor and the criterion. This index compares the standard error of criterion scores predicted by means of the test to the standard error of chance estimates. The proportionate reduction of the standard error is taken as a measure of the value of the test.

The "coefficient of determination," r^2, is another index that is used to evaluate tests. This index expresses the ratio of predicted variance in the criterion to the total variance. Use of this index and the index of forecasting efficiency requires that the correlation be reasonably high (about .50) in order to conclude that the test is substantially beneficial. The index of forecasting efficiency describes a test with such validity as predicting only 13 per cent better than chance, while the coefficient of determination describes such a test as accounting for 25 per cent of the variance in the criterion.

The major variation on this approach is due to Brogden (1946). He demonstrated mathematically--through manipulation of the formulas for $\underline{r}$--that $\underline{r}$, not $\underline{E}$ or $\underline{r}^2$, is a direct measure of the proportion improvement over chance prediction afforded by a selection test. Thus, an $\underline{r}$ of .05 indicates that the test provides five per cent of the improvement over chance that a perfect test would provide; an $\underline{r}$ of .50, 50 per cent; an $\underline{r}$ of .95, 95 per cent. This means, if the correlational approach is valid, that the units on the $\underline{r}$ scale are equal in value to the institution using the test, a great departure from the implications of $\underline{E}$ and $\underline{r}^2$ that the units near 1.00 are much more important than the units near zero. (For example, $\underline{E}$ is .004 greater for an $\underline{r}$ of .10 than for an $\underline{r}$ of .05, while it is .12 greater for an $\underline{r}$ of .95 than for an $\underline{r}$ of .90. This implies that the units between .90 and .95 are 30 times as important to the institution as the units between .05 and .10.)

Subsequently, Brogden (1949) developed an index of selection test value that avoided some of the restrictive assumptions of $\underline{r}$, namely, normal distributions and linear regression. When the empirical data conform to these assumptions, Brogden's index theoretically equals $\underline{r}$. He also advocated use of utility scales as criteria in place of conventional measures of performance.

Chapter II is devoted to pointing out some of the limitations of the correlational approach. A method based on Brogden's approach is developed and presented in Chapter IV. It is called the "utility function" method.

The Decision-Theoretic Approach

Taylor and Russell (1939) took the first major step toward the decision-theoretic approach. They contended that the value of a test varies with the particular decision to be made, and that the problem is one of improving selection rather than of simply raising the correlation of a test with some criterion measure. They showed that considerable benefit can be obtained from tests with rather low validity. Benefit was defined as the difference between the proportion of employees likely to be "satisfactory" before and after selection by means of the test. This difference was as much dependent on the a priori probability (commonly called the base rate) and the selection ratio as it was upon the validity coefficient. (This is demonstrated in Chapter II.)

The next major advance in this approach came 18 years later with the publication of the monograph Psychological Tests and Personnel Decisions by Cronbach and Gleser (1957). Cronbach and Gleser took the position that the ultimate purpose of any personnel testing program is to assist in making decisions in regard to what should be done with an individual, and that the soundest approach to evaluating a test or testing program is through determining the benefits which accrue to the institution or individual as a result of the decisions which have been made. These writers used the concept of "utility" as a measure of test value and defined it as the benefits which accrue from a set of decisions less the total costs which are incurred in the decision-making process. Thus, this approach is a pragmatic one stressing the consequences of direct action (selection decisions) instead of abstract standards of predictive efficiency.

The most formidable and complex aspect of carrying out this approach in practice is quantifying the relative utility of decisions outcomes. Cronbach and Gleser (1957) make no contribution to the solution of this problem, other than pointing to it and discussing its relevance. However, there is an extensive history of value measurement and psychological scaling which is directly applicable. The present study attempts to draw on this knowledge for a solution of the test evaluation problem.

It should be noted that decision theory did not introduce the problem of values into the decision process and hence into personnel selection. It does, however, make it explicit. Value systems have always entered into decisions, but they were not heretofore clearly recognized or systematically handled.

Plan of the Study

Chapter II is devoted to demonstrating some of the limitations of the correlational approach for evaluating selection tests. Then in Chapter III personnel selection on the basis of psychological tests is presented in statistical decision theory terms. It is shown that this theory treats the base rate, selection ratio and institutional gains and losses explicitly and systematically. This formulation of selection test theory, unlike the Cronbach and Gleser one, is restricted to the dichotomous (or dichotomized) criterion case and does not rely on the correlation coefficient as an index of association between the test and the criterion.

Two new indices for evaluating selection tests are developed in Chapter IV. One is based on statistical decision theory as presented in Chapter III and the other is based on Brogden's approach (1949).

The next two chapters, V and VI, deal with utilities and ways to measure them. Two psychological scaling methods are described and applied in an empirical situation. A way to determine payoff matrices given these scale values is presented. This method is applied to the scale values and the final payoff matrices are determined.

An empirical tryout of the new indices is reported in Chapter VII. Selection test scores and final grades were obtained for large samples of students in U. S. Navy technical schools. The index values, as well as <u>r</u> and <u>E</u>, are presented and compared in terms of their indications of the predictive efficiency and utility of the selection tests.

CHAPTER II

LIMITATIONS OF THE CORRELATIONAL APPROACH

There can be no doubt that validity coefficients dominate the test evaluation scene. Of the 426 abstracts in the Handbook of Employee Selection (1950), 236 use a validity coefficient as the sole measure of test value. Manuals of published tests rarely report anything on test value except validity coefficients. Only about one-half of the reviews of aptitude tests in the Fifth Mental Measurement Yearbook (1959) cite any evidence of test value other than validity coefficients. Of the 32 abstracts in the "Validity Information Exchange" of Personnel Psychology in 1959, not a single one reported any numerical analysis indicating test value except validity coefficients.

The inappropriateness of validity coefficients as selection test evaluation indices is due to the following four limitations. (In each case the statistical assumptions underlying validity coefficients, namely, normality, linearity, and homoscedasticity are granted. Since both of the special product-moment correlation coefficients recommended under these assumptions--r_b and r_{tet}--are approximations to a Pearson r and are generally equivalent to it when these assumptions are true [Guilford, 1956, pp. 297-310], the limitations apply to them as well. The question as to whether the limitations also apply to phi is not raised because point distributions, or "genuine dichotomies," are not discussed.)

Validity Coefficients are Independent of the Selection Ratio

The selection ratio is the proportion of applicants (or population tested) to be accepted. It may be any proportion between zero and 1.00. The validity coefficient is independent of the selection ratio but test value is not. Consider Table 1 where the entries are the proportion of accepted applicants who are satisfactory in terms of job proficiency. These entries can be compared with the a priori probability .50 which is the proportion that would have been satisfactory had selection been random. The variation in each row shows variation in test value which is not accounted for by the correlation between test score and job proficiency. Take for instance the row pertaining to an r of .50; if the selection ratio is .05, 88 per cent will be satisfactory, a sizeable increase over the a priori probability; if the selection ratio is .95, 52 per cent will be satisfactory, a very slight improvement over the a priori probability. The correlation may not adequately indicate the value of the test in any specific situation. It can be seen from the table that a test with almost any validity may or may not be of much value depending upon the selection ratio.

Validity Coefficients are Independent of the A Priori Probability

The a priori probability is the proportion who will be satisfactory if selection is random. Test value is very much dependent upon the a priori probability but validity coefficients are not. Table 2 will clarify this. As in Table 1 the entries are the proportions of accepted

TABLE 1

THE PROPORTION WHO WILL BE SATISFACTORY AMONG THOSE SELECTED, WHEN THE A PRIORI PROBABILITY IS .50 (FROM TAYLOR AND RUSSELL, 1939)

	Selection Ratio										
r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.05	.54	.54	.53	.52	.52	.52	.51	.51	.51	.50	.50
.10	.58	.57	.56	.55	.54	.53	.53	.52	.51	.51	.50
.15	.63	.61	.58	.57	.56	.55	.54	.53	.52	.51	.51
.20	.67	.64	.61	.59	.58	.56	.55	.54	.53	.52	.51
.25	.70	.67	.64	.62	.60	.58	.56	.55	.54	.52	.51
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52	.51
.35	.78	.74	.70	.66	.64	.61	.59	.57	.55	.53	.51
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53	.52
.45	.85	.81	.75	.71	.68	.65	.62	.59	.56	.53	.52
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54	.52
.55	.91	.87	.81	.76	.72	.69	.65	.61	.58	.54	.52
.60	.94	.90	.84	.79	.73	.70	.66	.62	.59	.54	.52
.65	.96	.92	.87	.82	.77	.73	.68	.64	.59	.55	.52
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55	.53
.75	.99	.97	.92	.87	.82	.77	.72	.66	.61	.55	.53
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55	.53
.85	1.00	.99	.97	.94	.88	.82	.76	.69	.62	.55	.53
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56	.53
.95	1.00	1.00	1.00	.99	.96	.90	.81	.71	.63	.56	.53
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56	.53

TABLE 2

THE PROPORTION WHO WILL BE SATISFACTORY AMONG THOSE SELECTED, WHEN THE SELECTION RATIO IS .50 (FROM TAYLOR AND RUSSELL, 1939)

	A Priori Probability										
r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.05	.05	.11	.21	.31	.42	.52	.62	.71	.81	.91	.95
.10	.06	.11	.22	.33	.43	.53	.63	.73	.82	.91	.96
.15	.06	.12	.23	.34	.45	.55	.65	.74	.83	.92	.96
.20	.07	.13	.25	.36	.46	.56	.66	.76	.84	.93	.97
.25	.07	.13	.26	.37	.48	.58	.68	.77	.86	.93	.97
.30	.07	.14	.27	.38	.49	.60	.69	.78	.87	.94	.97
.35	.08	.15	.28	.40	.51	.61	.71	.80	.89	.95	.98
.40	.08	.16	.29	.41	.53	.63	.73	.81	.89	.95	.98
.45	.08	.16	.30	.43	.54	.65	.74	.83	.90	.96	.98
.50	.09	.17	.31	.44	.56	.67	.76	.84	.91	.97	.99
.55	.09	.17	.32	.46	.58	.69	.78	.86	.92	.97	.99
.60	.09	.18	.34	.47	.60	.70	.80	.87	.94	.98	.99
.65	.10	.18	.35	.49	.62	.73	.82	.89	.95	.98	1.00
.70	.10	.19	.36	.51	.64	.75	.84	.91	.96	.99	1.00
.75	.10	.19	.37	.52	.66	.77	.86	.92	.97	.99	1.00
.80	.10	.20	.38	.54	.68	.80	.88	.94	.98	1.00	1.00
.85	.10	.20	.39	.56	.71	.82	.91	.96	.99	1.00	1.00
.90	.10	.20	.40	.58	.74	.86	.94	.98	1.00	1.00	1.00
.95	.10	.20	.40	.60	.77	.90	.97	.99	1.00	1.00	1.00
1.00	.10	.20	.40	.60	.80	1.00	1.00	1.00	1.00	1.00	1.00

applicants who are satisfactory in terms of job proficiency. Comparison of each entry with the appropriate a priori probability, i.e., the one that heads the column in which the entry is located, provides a meaningful indication of test value. The difference between the a priori probability and an entry is the improvement over chance which the predictor makes possible. The variation in these differences within any row is the variation in test value that is not accounted for by the validity coefficient which heads that row. For example, the differences between the a priori probabilities and the entries in the row pertaining to an *r* of .50 are .04, .07, .11, .14, .16, .17, .16, .14, .11, .07, .04. These differences for an *r* of .80 are .05, .10, .18, .24, .28, .30, .28, .24, .18, .10, .05.

We may conclude therefore, that the validity coefficient may not adequately represent the value of a test in a specific situation. Even a very high correlation is not very good evidence that the test is worth much. A test that correlates .90 with a criterion may be worth no more than a test that correlates .30 with a criterion: when the first criterion has an a priori probability of .10 or .90 and the second criterion has an a priori probability of .50--the differences between the a priori probabilities and the corresponding entries in the table are equal.

All Errors of Measurement Attenuate the Validity Coefficient

When all observations in the criterion-test plot fall in a straight line, the correlation is perfect, i.e., *r* = 1.00. Any deviations from a straight line result in an *r* less than 1.00. Such deviations are said to "attenuate" *r*. Therefore, when *r*, or any correlation coefficient derived from *r* such as the biserial *r* and the tetrachoric *r*, is used as an evaluation index, the assumption is implicit that all deviations from the line representing perfect correlation are important. In other words all such deviations are assumed to have practical significance. It can be argued, however, that only deviations which affect the decision for which the test is used should attenuate the evaluation index.

When a psychological test is used as an aid in making decisions, the most common practice is to set a cutoff on the test scale and make one decision about persons who receive a score above that point and the complementary decision about persons who receive a score below that point. In the personnel selection situation the decisions are to accept or to reject the persons for the assignment. Such a situation is depicted in Figure 1. The cutoff is labeled x_c. The line passing

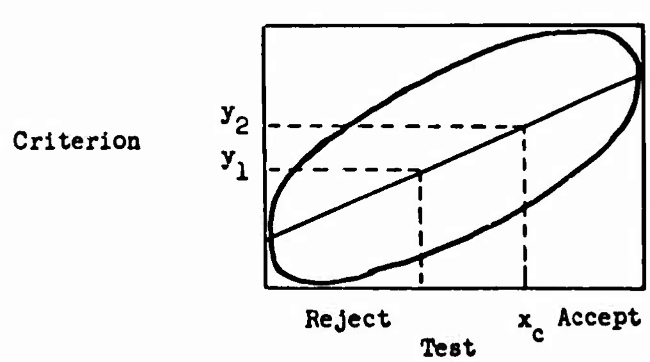

Fig. 1.--An exemplary scatter plot showing the regression line and the cutoff used in making decisions.

through the plot is the regression line, the line of best fit in a least-squares sense (Guilford, 1956, p. 366). If a person whose score on the test exceeds x_c receives a score located at y_1 on the criterion, the test could be said to have made an erroneous decision since, had the test predicted perfectly, this person would have been rejected. However, if this "accepted" person received a criterion score above y_2, regardless of which one, the decision based on the test must be considered correct. Similarly, the decision to reject a person must be considered correct if his criterion score is below y_2.

The establishment of, and adherence to, a cutoff divides the scatter plot into four areas shown in Figure 2. Deviations from the regression line in areas B and C are not errors and should not attenuate

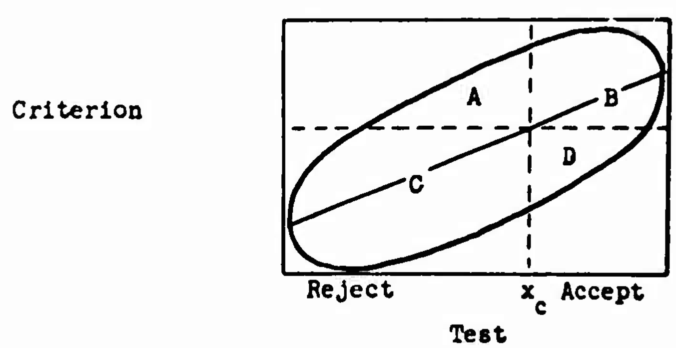

Fig. 2.--A scatter plot showing the four decision-related areas determined by the cutoff and the regression line.

the evaluation index if it is to be taken as an estimate of the value of the test in this decision situation. Only deviations which lead to an erroneous decision should be considered errors. These are observations which fall in areas A and D. Validity coefficients, of course, consider every observation that falls off the regression line as an error regardless of its importance to the decision.

Furthermore, the size of a deviation from the regression line in areas B and C is irrelevant. All observations in each of these cells should receive equal weight in the evaluation index since they are all equally correct--a perfect test would have led to the same decision in every case and to, therefore, the same consequences. This is not true of the validity coefficient, which weights observations in proportion to the size of their deviations from the regression line. It seems reasonable to contend that differential weighting within these areas is illogical when attempting to determine the value of a test for a dichotomous decision.

Validity Coefficients do not Adequately Reflect Institutional Gains and Losses

A validity coefficient in selection testing is an index of strength of predictive association between a selection test and a criterion (usually some measure of performance). As such, the only link with institutional gains and losses is through the criterion. Implicit in the use of r as an evaluation index is the assumption that the utility function of the criterion is linear, i.e., that equal increments of the criterion represent equal increments of utility or value to the institution using the test. This assumption is rarely tested with quantitative research. In fact, it is rarely mentioned in the psychometrics literature.

Following the logic of the previous section, a more reasonable assumption in general for selection tests would be that the utility function is stepwise about the point on the criterion corresponding to the test cutoff. Consideration of this point is what usually leads to the choice of the cutoff. It seems reasonable to expect the criterion units around this point to be more important to the institution than those far above or below this point.

Actually, of course, the shape of the utility function of the criterion in an applied situation is an empirical question to be answered ideally through research. In the absence of such research the most reasonable assumption should be stated and an evaluation index used which does not violate that assumption. In selection test evaluation it would seem that any evaluation index based on product-moment correlation theory should be avoided.

Another point mentioned in the previous section is that observations which fall off the regression line are weighted by the validity coefficient in proportion to their distance from the regression line. Institutional gains and losses are not expressly taken into account. The two extreme types of deviations are commonly called false positives and false negatives. (In subsequent chapters these are called erroneous acceptees and erroneous rejectees.) The implicit assumption in correlational analysis is that these are equally costly to the institution using the test. Whether or not they are equally costly is an empirical question. Their actual cost to the institution should be determined through research.

In this chapter some of the inadequacies of the conventional approach to selection test evaluation have been discussed showing that a new approach is needed and that a more adequate approach should handle the following factors:

(1) selection ratio,

(2) a priori probability,

(3) institutional gains and losses.

The next chapter presents the theoretical foundation of an approach based on statistical decision theory which handles these factors explicitly and systematically.

CHAPTER III

SELECTION TESTS AND STATISTICAL DECISION THEORY

The monograph by Cronbach and Gleser (1957) was the first and most direct, large scale restatement of test evaluation theory in the decision-theoretic framework. The present chapter outlines a somewhat simpler, more straightforward approach to what Cronbach and Gleser call "selection decisions with single-stage testing," which, unlike their approach, does not rely on correlation coefficients. It is restricted to situations in which the criterion is dichotomous (or dichotomized) and the test score is continuous.

Statistical decision theory specifies the optimum decision in a situation where one must choose between two alternative statistical hypotheses on the basis of an observed event. In particular, it specifies the optimum cutoff, along the continuum on which the observed events are arranged, as a function of (a) the a priori probabilities of the two hypotheses, (b) the values and costs associated with the various decision outcomes, and (c) the amount of overlap of the distributions that correspond to the hypotheses. See especially Chernoff and Moses (1959), Good (1962), Marschak (1954), and Swetts et al. (1961).

In applied psychology, selection tests are most often used to make a simple yes-no decision in terms of such things as hiring, promotion, training, etc. A particular dichotomous decision represents predictions (or hypotheses) based on a test score. In Figure 3 test score is labeled $\underline{x}$ and plotted on the abscissa. The left-hand distribution, labeled $f_F(x)$, is the probability density function of $\underline{x}$ given a person who would "fail." The right-hand distribution is the probability density function of $\underline{x}$ given a person who would "succeed." (Probability density functions are used, rather than probability functions, since $\underline{x}$ is assumed to be continuous.) Although the distributions appear to be normal and equally variant, the selection test model presented below assumes neither.

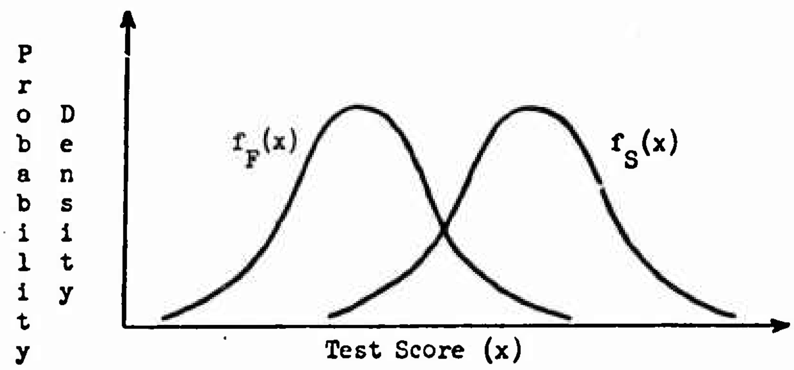

Fig. 3.--The probability density function of Fail and Succeed.

The basic decision is whether a given test score arises from one distribution or the other, or, equivalently, the relative probabilities that a person obtaining that score will succeed or fail. It is desirable to establish a standard, a cutoff x_c on the continuum of test scores, to which any given score x_i can be related. If it is found for the i-th test score, x_i, that $x_i > x_c$, the decision is to "accept"; if $x_i < x_c$, the decision is to "reject."

In the language of statistical decision theory a subset of all the scores, namely a Critical Region A (accept), is chosen such that a test score in this subset leads to acceptance of the Hypothesis S, to the prediction that the person will succeed. All other scores are in the complementary subset R (reject); these lead to rejection of the Hypothesis S, or, equivalently, to the acceptance of the Hypothesis F, to predict the person will fail. The Critical Region A, with reference to Figure 3, consists of the values of $\underline{x}$ to the right of some cutoff x_c.

The decision outcome may be a correct acceptance (A,S--the joint occurrence of a score in Region A and success), a correct rejection (R,F), an erroneous rejection (R,S), or an erroneous acceptance (A,F). If the a priori probability of a success and the parameters of the distributions of Figure 3 are fixed, the choice of a cutoff value x_c completely determines the probability of each of these outcomes.

Clearly, the four probabilities are interdependent. For example, an increase in the probability of a correct acceptance, P(A,S), can be achieved only by accepting an increase in the probability of an erroneous acceptance, P(A,F), and decreases in the other probabilities, P(R,S) and P(R,F). Thus, a given cutoff yields a particular balance among the probabilities of the four possible outcomes; conversely, the balance desired in any instance will determine the optimum location of the cutoff. Now one may desire the balance that maximizes the expected value of decisions where the four possible outcomes have individual utilities. One may, however, desire a balance that maximizes some other quantity--i.e., a balance that is optimum according to some other definition of optimum--in which case a different cutoff will be appropriate. One may, for example, want to maximize P(A,S) while satisfying a restriction on P(A,F), as one typically does when as an experimenter one assumes an .05 or .01 level of confidence. Alternately, one may want to maximize the number of correct decisions.

The manner of specifying the optimum cutoff will be illustrated for just one of these definitions of optimum, namely, the maximization of the total expected value (or utility) of a decision in a situation where the four possible outcomes of a decision have individual utilities associated with them. The expected utility (EU) of a strategy is defined in statistical decision theory as the sum, over the potential outcomes of a decision, of the products of probability of outcome and the desirability (utility) of outcome:

$$EU = P(A,S)U_{A,S} + P(A,F)U_{A,F} + P(R,F)U_{R,F} + P(R,S)U_{R,S}.$$

In this equation $U_{A,S}$, $U_{A,F}$, $U_{R,F}$, $U_{R,S}$, are the utilities of a correct acceptance, an erroneous acceptance, a correct rejection, and an erroneous rejection, respectively. For any observed value, x_i, the expected utility of the decision to accept is:

$$EU_A = P(S|x_i)U_{A,S} + P(F|x_i)U_{A,F},$$

where $P(S|x_i)$ is the probability of a "success" conditional upon, or given, x_i; $P(F|x_i)$ is the probability of a "fail," given x_i. Similarly, the expected utility of the decision to reject is given by

$$EU_R = P(F|x_i)U_{R,F} + P(S|x_i)U_{R,S}.$$

In statistical decision theory the optimum cutoff is specified in terms of the likelihood ratio:

$$\lambda(x) = \frac{f_S(x)}{f_F(x)},$$

which is the relative probability that a person obtaining score _x_ will succeed or fail. It will be shown that the optimum cutoff can be specified by some value of $\lambda(x)$, provided that (1) $\lambda(x)$ is monotonic increasing with x and (2) the utilities of correct decisions are greater than the utilities of the complimentary erroneous decisions, i.e., $U_{A,S} > U_{R,S}$ and $U_{R,F} > U_{A,F}$.

Given these conditions, EU_A will equal EU_R at the optimum cutoff, x_c, since it is the point on the test score scale where it makes no difference whether the testee is accepted or rejected--the expected payoff is the same in either case. Thus

$$P(S|x_i)U_{A,S} + P(F|x_i)U_{A,F} = P(F|x_i)U_{R,F} + P(S|x_i)U_{R,S} \tag{1}$$

or

$$P(S|x_i)(U_{A,S} - U_{R,S}) = P(F|x_i)(U_{R,F} - U_{A,F}). \tag{2}$$

Cross multiplying yields

$$\frac{P(S|x_i)}{P(F|x_i)} = \frac{(U_{R,F} - U_{A,F})}{(U_{A,S} - U_{R,S})}. \tag{3}$$

Now according to Bayes' rule (Good, 1962), the likelihood ratio, $\lambda(x)$, for any x_i can be expressed as

$$\lambda(x_i) = \frac{P(F)P(S|x_i)}{P(S)P(F|x_i)}. \tag{4}$$

Multiplying both sides of Equation (3) by P(F)/P(S) yields

$$\frac{P(F)P(S|x_i)}{P(S)P(F|x_i)} = \frac{P(F)(U_{R,F} - U_{A,F})}{P(S)(U_{A,S} - U_{R,S})} . \qquad (5)$$

The left-hand term of Equation (5) is the likelihood ratio given in Equation (4). Thus, the likelihood ratio at the optimum cutoff has been shown to be equal to the right-hand term in Equation (5). That is, it is the point on the $\underline{x}$ continuum where Equation (5) is true.

The optimum cutoff can be specified by some value B of $\lambda(x)$. This value can now be given as

$$B = \frac{P(F)(U_{R,F} - U_{A,F})}{P(S)(U_{A,S} - U_{R,S})} , \qquad (6)$$

since, when $\lambda(x) > B$, $EU_A > EU_R$, and when $\lambda(x) < B$, $EU_A < EU_R$. This can be seen by noting that when $\lambda(x) > B$ this inequality will also be true of Equations (5), (3), (2), and (1); consequently $EU_A > EU_R$. Similarly, when $\lambda(x) < B$ this inequality will be true of the same equations, and $EU_A < EU_R$. The decision should therefore be to "accept" whenever $\lambda(x) > B$ and to "reject" whenever $\lambda(x) < B$. The former will be true only when $x > x_c$ and the latter only when $x < x_c$, provided that $\lambda(x)$ is monotonic increasing with $\underline{x}$, $U_{A,S} > U_{R,S}$, and $U_{R,F} > U_{A,F}$. Thus the Critical Region A lies to the right of x_c and the Critical Region R lies to the left of x_c.

This constitutes the model of test-based selection decisions from the standpoint of statistical decision theory. The cutoff (and therefore, the selection ratio), a priori probability and institutional gains and losses are central factors. An evaluation index based on this model is presented in the next chapter. Chapters V and VI deal with utilities and ways to measure them.

CHAPTER IV

TWO NEW METHODS FOR EVALUATING SELECTION TESTS

In this chapter an index for evaluating selection tests which is based on the model presented in the previous chapter is developed. It will be seen that no index of association is needed because the evaluation index represents a direct measure of the improvement over chance prediction provided by the test. In the final section of this chapter is presented an index based on the method developed by Brogden (1949) which also purports to indicate the utility of selection tests.

Decision-Theoretic Method

The starting point of this method is a payoff matrix. When a cutoff on the test is used and the outcomes to be predicted form a dichotomy, the payoff matrix is as shown in Figure 4; where U_1, U_2, U_3, and U_4 are utilities which correspond to erroneous rejection, correct acceptance, correct rejection, and erroneous acceptance, respectively. (See Chapter VI for a thorough explanation of payoff matrices.)

Criterion (Job A)		Reject	Accept
	Succeed	U_1	U_2
	Fail	U_3	U_4
		Decision	

Fig. 4.--The standard payoff matrix.

Assume that 100 persons, selected at random, have been assigned to job A. The utility equation for an obtained table is

$$U = n_1U_1 + n_2U_2 + n_3U_3 + n_4U_4, \tag{1}$$

where the n's are the frequencies in the corresponding cells of the contingency table shown in Figure 5. To estimate the utility of a

		1 - q	q	
Criterion (Job A)	Succeed	n_1	n_2	p
	Fail	n_3	n_4	1 - p
		Low	High	
		Test		

Fig. 5.--The standard 2 X 2 contingency table.

test to the decision-making process, U must be compared with the one that would result with a test of zero utility, i.e., one providing only chance prediction, U_c. When the observations in the contingency table are randomly distributed, each cell frequency is the product of the corresponding marginal probabilities and N. $(N = n_1 + n_2 + n_3 + n_4)$. Therefore

$$U_c = (p - pq)NU_1 + pqNU_2 + (1 + pq - p - q)NU_3 + (q - pq)NU_4, \tag{2}$$

where p is the a priori probability and q is the selection ratio as shown in Figure 5. Then, the utility of the test is given by the difference between U and U_c:

$$U_T = U - U_c \tag{3}$$

This procedure can be simplified and made to fit the usual test evaluation situation where n_1 and n_3 are not known. It can be shown (see Appendix A) that U_T is independent of the addition of any constant (positive or negative) to the values of both entries in a row of the payoff matrix. Since only the individuals above the cutoff on the test, the accepted group, are available to the test evaluater, the most useful payoff matrix is the one shown in Figure 6.

Criterion (job A)		Decision: Reject	Decision: Accept
	Succeed	0	$U_2 - U_1$
	Fail	0	$U_4 - U_3$

Fig. 6.--The modified payoff matrix obtained by subtracting U_1 from the first row and U_3 from the second row.

Then,

$$U = n_2(U_2 - U_1) + n_4(U_4 - U_3) \qquad (4)$$

and

$$U_c = pqN(U_2 - U_1) + (q - pq)N(U_4 - U_3). \qquad (5)$$

Since N and q are unknowns, substitute for q its equivalent, $(n_2 + n_4)/N$:

$$U_c = p\left(\frac{n_2 + n_4}{N}\right)N(U_2 - U_1) + (1 - p)\left(\frac{n_2 + n_4}{N}\right)N(U_4 - U_3). \qquad (6)$$

Now N cancels and U_c becomes

$$U_c = p(n_2 + n_4)(U_2 - U_1) + (1 - p)(n_2 + n_4)(U_4 - U_3). \qquad (7)$$

Again, the difference between U and U_c equals the gain in utility due to the test:

$$U_T = U - U_c$$

This U_T is equal to the one obtained prior to changing the payoff matrix. Appendix A presents the mathematical proof.

Example: Assume that 100 men were assigned to electronic training and that, after training, the graduates and fails are distributed as in Table 3.

TABLE 3

A HYPOTHETICAL CONTINGENCY TABLE

		Test: Low	Test: High	
Criterion	Graduate	20	60	p = .8
	Fail	15	5	1 - p = .2

Assume further that the consequences of the four decision-outcome combinations have been considered (see Chapter VI) and the payoff matrix shown in Table 4 has been determined.

TABLE 4

A HYPOTHETICAL PAYOFF MATRIX

		Decision: Reject	Decision: Accept
Criterion	Graduate	-8	10
	Fail	12	-6

The U equation for this example is

$$U = 20(-8) + 60(10) + 15(12) + 5(-6)$$
$$= -160 + 600 + 180 - 30$$
$$= 590.$$

The $\underline{U}$ equation for chance prediction is

$$U_c = (.8 - .52)100(-8) + (.52)100(10) + (1 + .52 - .8 - .65)$$
$$100(12) + (.65 - .52)100(-6)$$
$$= 28(-8) + 52(10) + 7(12) + 13(-6)$$
$$= 302.$$

The utility of the test is

$$U_T = 590 - 302 = 288.$$

If the payoff matrix is simplified as shown above it becomes the one presented in Table 5; then,

$$U = 60(18) + 5(-18) = 990$$

and

$$U_c = .8(60 + 5)(18) + .2(60 + 5)(-18)$$
$$= 52(18) + 13(-18)$$
$$= 702.$$

The utility is the same as before:

$$U_T = 990 - 702 = 288.$$

An assumption explicit in this method is that the cutoff has been set at the best possible point on the test. If an inflexible quota must be filled this assumption is of no consequence. However, many times it is of value to determine the best possible cutoff, i.e., one that balances the positive and negative utilities of correct and

TABLE 5

A MODIFIED VERSION OF A HYPOTHETICAL PAYOFF MATRIX

Criterion		Reject	Accept
	Graduate	0	18
	Fail	0	-18
		Decision	

erroneous decisions. This point can be easily determined if a payoff matrix is available. It has been shown in Chapter III to be the point on the test where

$$\lambda(x) = \frac{P(F)(U_{R,F} - U_{A,F})}{P(S)(U_{A,S} - U_{R,S})}, \tag{8}$$

where $\lambda(x)$ is the likelihood ratio $f_S(x)/f_F(x)$. In the symbolism of contingency tables and payoff matrices, the right-hand term of Equation (8) is

$$\frac{(1 - p)(U_3 - U_4)}{p(U_2 - U_1)}.$$

The test will be of greatest utility if the cutoff is set at the point where

$$\lambda(x) = \frac{(1 - p)(U_3 - U_4)}{p(U_2 - U_1)}, \tag{9}$$

or, according to Equation (4) in Chapter III, where

$$\frac{(1 - p)P(S|x_1)}{pP(F|x_1)} = \frac{(1 - p)(U_3 - U_4)}{p(U_2 - U_1)}, \tag{10}$$

which can be reduced to

$$\frac{P(S|x_i)}{P(F|x_i)} = \frac{U_3 - U_4}{U_2 - U_1} . \tag{11}$$

Utility Function Method

This method is essentially the one developed by Brogden (1949). He was concerned, however, with the case in which the test is dichotomous and the criterion is continuous. The method is described here for the case in which both variables are in dichotomous form.

The criterion is translated into utility terms and the "gain" per man selected is computed. Consider Figure 5 in the preceding section where the observations are a random sample of size N ($N = n_1 + n_2 + n_3 + n_4$) from the population in which the test is to be used. All N persons have been assigned to job A. Test scores have been obtained for all N persons prior to their assignment to job A. Criterion scores, succeed and fail, have been assigned on the basis of performance in job A and translated into utility terms. An individual's criterion score is his utility in job A. These utility values are hereafter labeled U_S and U_F.

From the above definitions the following statistics can be determined:

$$M_U = \frac{(n_1 + n_2)U_S + (n_3 + n_4)U_F}{N} \tag{12}$$

Equation (12) can be interpreted as "the mean utility for a random sample of individuals assigned to job A."

$$U = \frac{n_2 U_S + n_4 U_F}{n_2 + n_4} \tag{13}$$

Equation (13) can be interpreted as "the average utility for the subgroup of a random sample of individuals who are high on the test when they are assigned to job A."

$$G_U = U - M_U \tag{14}$$

The value "G_U" defined in Equation (14) is the gain in utility which would be realized, on the average, by assigning individuals to job A on the basis of the test, rather than at random.

Example: Assume that 100 Navy recruits were assigned at random to electronic training and that, after training, the graduates and fails (non-graduates) are distributed as in Table 3 in the previous section. Assume further that a graduate is worth 100 utiles (the unit of measurement on the utility scale) to the Navy, and a fail is worth 40 utiles to the Navy. The total utility of these men to the Navy is easily determined. There are 80 graduates, worth 100 utiles each, or 8,000 utiles altogether. There are 20 fails worth 40 utiles each, or 800 utiles altogether. Thus, the total utility for the group is 8,800 utiles. The average utility for the men assigned to electronics training is

$$M_U = \frac{8800}{100} = 88.$$

For men high on the test, the average utility is similarly determined to be

$$U = \frac{(60)(100) + (5)(40)}{65} = 95.38.$$

Then the gain per man is

$$G_U = 95.38 - 88 = 7.38.$$

The conclusion would be that, provided the manpower pool is large enough, the Navy will be 7.38 utiles ahead, on the average, for each man assigned using the test. This figure should of course be reduced by the cost of testing. This cost will be ignored here because it is negligible per man in the Navy setting. (Testing takes one day out of a recruit's schedule, and four men administer a test battery to 500 recruits per day.)

Since $\underline{N}$, n_1, and n_3 are not known when the test to be evaluated has been operational for some time, it will help to express the equation for M_U in terms of the a priori probability, $\underline{p}$, estimated from previous research. An equivalent equation is

$$M = pU_S + (1 - p)U_F \tag{15}$$

Unlike the method presented in the previous section, this method does not consider the cost of rejecting a person who would have succeeded or the value of correctly rejecting a fail. G_U only reflects the gain <u>per selectee</u> over chance prediction. Therefore, it is to be expected that G_U will provide a lower estimate of the utility of selection tests than will U_T. Data bearing on this point will be found in Chapter VII.

G_U and U_T can be compared directly (mathematically) by making assumptions regarding the relative size of the utilities in the two methods. The most obvious is that $U_S = U_2 - U_1$ and $U_F = U_4 - U_3$.

When this is true $U_T = (n_2 + n_4)G_U$. However, these assumptions are very restrictive and will be true only rarely. They are not true in the empirical situation under study. Since $U_S = U_2$, U_S can equal $U_2 - U_1$ only when U_1 is zero. Also, since one assumption underlying the decision-theoretic approach is that $U_3 > U_4$ (see Chapter III), U_F can equal $U_4 - U_3$ only when U_F is negative, which is not true in the empirical situation under investigation.

Another method that might appeal to some readers is to weight the cell frequencies by the corresponding utilities and compute the phi coefficient. This method would have the following drawbacks: (1) n_1 and n_3 are often not known, (2) determining U_3 and U_4 would probably be more difficult than determining $U_4 - U_3$ (see the following chapter), and (3) the resultant coefficient would seem (to the writer) to be very difficult to interpret.

CHAPTER V

MEASUREMENT OF VALUES INHERENT IN TEST EVALUATION

Both of the approaches presented in the preceding chapter--one using a payoff matrix and the other a converted criterion scale--require quantitative measurement of value. The relative values of the four decision-outcome combinations must be determined in the first approach. The value of a satisfactory assignee relative to an unsatisfactory one must be determined in the second approach. In both cases, of central importance is the value of obtaining a satisfactory person for the assignment--U_2 or U_S. It can be thought of as the need for a satisfactory assignee. This value can be made more meaningful to the institution using the test by scaling the job areas on need--the need for a satisfactory assignee. The relative need for additional satisfactory persons in the job areas can in this way be determined and expressed quantitatively. Possible ways of scaling the job areas on need are described. How the criterion can be converted once this scale is obtained is shown.

The specific situation in terms of which these methods were explored, is that of recruit classification in the U. S. Navy. Selection tests are used on which cutoffs are established. If a recruit receives a score above the cutoff he may if he wishes go to the school for which the test is a selector (subject to quota restrictions); if he receives a score below the cutoff he will not be sent to that school. The criterion against which selection tests are currently validated is school grade. The methods described below are presented in terms of the dichotomous criterion, graduate-fail, which is based on school grade. The continuum on which job areas were scaled is therefore the utility of school graduates to the operational Navy, or, the need for school graduates in the corresponding job areas.

It might be worth mentioning at this point that a side benefit of this scaling process is that the scale values are vitally needed for optimal classification of recruits to schools and hence to job areas. Optimal classification is not possible without a measure of need across job areas. The same is true regarding job applicants in other applied situations.

Scaling Job Areas on Need

Two methods were used, one "indirect" method (probability comparison) which involves inferring values from choices made by judges, and one "direct" method (magnitude estimation) which requires each judge to estimate need in each job category. The methods are designed for different types of judges, namely, classification interviewers and area personnel planners. The indirect method was developed by the author of this study. He knows of no similar method in the scaling literature.

Probability Comparison

Eleven classification interviewers were asked to indicate how they would classify imaginary recruits with certain probabilities of success in Navy schools. A questionnaire (see Appendix C) was constructed containing items like the following for each pair of schools:

> To which school would you assign a recruit if you think his chances of success are

	School A		School B
(a)	80% ______	and	60% ______
(b)	80% ______	and	70% ______
(c)	80% ______	and	80% ______
(d)	80% ______	and	90% ______
(e)	80% ______	and	95% ______

In this way each respondent's indifference point for each pair of schools was determined. This is a point in the probability space where the respondent is indifferent as to the assignment of recruits to one school or the other. Its coordinates are assumed to be the midpoints of the intervals where the respondent's marks change columns. (The items should be so constructed to ensure that a crossover always occurs.)

Establishment of the indifference point for a pair of schools leads to the following equation:

$$pG_A + (1 - p)F_A = qG_B + (1 - q)F_B.$$

Here $\underline{p}$ is the probability that the recruit would graduate in school A, $\underline{q}$ is the graduation probability in School B that leads to indifference,

G_A and G_B are the subjective values of the recruit graduating in schools A and B, respectively, and F_A and F_B are the values of the recruit failing in schools A and B, respectively.

One restraint must be placed on the above equation in order to solve for G_A and G_B. It was assumed that $F_A = F_B$, i.e., that in making his choices the respondent considered the two events, failing in school A and failing in school B, equally bad. One point on the scale was established by making the following restriction:

$$F_A = F_B = 0.$$

The other arbitrary point on the scale was set by choosing a value for either G_A or G_B. If G_A is set equal to 100 the equation becomes

$$p(100) = qG_B.$$

If the point of indifference chosen by the respondent is defined by the probabilities .8 and .6, G_B can be determined thusly:

$$.8(100) = .6G_B.$$

$$G_B = 80/.6 = 133.$$

With 10 schools there are 45 possible pairs but only nine are necessary to scale them. Additional ones were presented in order to obtain stable scale values. The scale values of all can be computed once a single one is arbitrarily set.

The scaling questionnaire is presented in Appendix C. The percentage value for the school presented on the left in each question was set by the writer at the level that he thought would seem reasonable to the respondents. The percentage values for the school on the right were chosen to constitute adequate range of choices around the percentage value for the school on the left to ensure that a crossover would occur. As nearly as possible, the schools appear on the left an equal number of times.

There are 24 questions in the questionnaire, each with a unique pair of schools. Pairs were chosen in the following way: the schools were subjectively ranked by the writer in terms of the need for additional men in the job areas corresponding to the schools; the nine adjacent pairs were used; a wide variety of more divergent pairs were chosen in such a way that the schools appear roughly the same number of times.

Eleven indifference points were obtained for each question--one from each respondent. In the rare instance where all the marks were in one column, the indifference point was assumed to be the point represented at the low end of the second column.

The data are presented in Table 6. The mean indifference point for each question in the questionnaire is presented as is the average deviation. (Standard deviations were not used because the distributions are somewhat truncated at the upper end and a few extreme scores occur at the other end.) The indifference points in column (4) are not directly comparable since they represent responses made relative to various dissimilar chances of success: those presented in column (3). The indifference points were made comparable by dividing the chance of

TABLE 6

MEAN INDIFFERENCE POINTS, AVERAGE DEVIATIONS, AND THE RATIOS USED IN CALCULATING THE SCALE VALUES FOR THE PROBABILITY COMPARISON SCALING METHOD

Question (1)	School Pair (2)	Chance of Success in First School (3)	Mean Indifference Point (4)	Average Deviation (5)	Preference Ratio* (6)
1	RM--EN	70	87	5.91	.805
2	PC--YN	90	79	8.90	1.139
3	RM--HM	70	85	4.18	.824
4	DK--HM	80	70	6.82	1.114
5	MM--SO	90	68	6.20	1.324
6	SK--EN	90	85	3.54	1.059
7	EN--MM	80	76	4.00	1.053
8	PC--EN	90	81	7.09	1.111
9	ET--RM	70	67	4.54	1.045
10	SO--ET	70	76	6.64	.921
11	HM--SK	80	85	2.00	.941
12	EN--DK	80	84	1.73	.952
13	SK--PC	90	91	3.27	.989
14	DK--ET	80	66	5.18	1.212
15	DK--YN	80	73	5.64	1.096
16	ET--HM	70	85	5.73	.824
17	MM--SK	90	90	4.64	1.000
18	HM--SO	80	67	4.82	1.194
19	YN--SK	80	84	4.45	.952
20	SO--RM	70	71	6.91	.985
21	HM--YN	80	86	4.64	.930
22	RM--DK	70	91	3.73	.769
23	YN--EN	80	83	5.82	.964
24	PC--MM	90	77	8.18	1.169

*The entry in column (3) divided by the entry in column (4).

success in the school presented on the left in each item by the mean indifference point for that item. These ratios are presented in column (6) of Table 6. The product of a ratio and the scale value of the first school yields the scale value of the second school. Using these ratios and an arbitrarily assigned number for the value of a graduate from one of the schools it is possible to compute scale values for all of the schools. This was done several times using different schools as the arbitrary base. These computations showed that graduates of SO school are worth more to the Navy than graduates of any other of the ten schools. For the computation of the final scale values the value of 100 was assigned to SO to represent the upper end of the need scale.

The order of computation was as follows: Scale values were first computed for schools in questions in which SO appeared on the left. Then, order of computation was dictated by the order in which these scale values were obtained. That is, as a scale value for a school was obtained, this value was used with the questions in which that school appeared on the left. This procedure was followed until all the questions had been used.

Table 7 contains the scale values obtained from the twenty-four questions. In each case the scale value was computed by multiplying the preference ratio--column (6) of Table 6--for a question by an already computed (or assigned, for SO) scale value. Table 7 also contains the mean scale values for the ten schools under study. These are the relative utilities of school graduates as determined by the probability comparison scaling method.

TABLE 7

SCALE VALUES AND RELATIVE UTILITIES OF SCHOOL GRADUATES AS SCALE VALUES OBTAINED THROUGH THE PROBABILITY COMPARISON SCALING METHOD

School	Scale Values from Individual Questions			Utilities (Mean Scale Values)
SO	(100.0)	90.6	110.5	100.4
RM	98.5	96.2		97.4
ET	92.1	91.8		92.0
MM	83.5	82.6		83.0
HM	75.9	81.2	84.4	80.5
YN	70.6	83.0	80.4	78.0
DK	75.8	75.5		75.6
EN	79.3	75.6	68.0	74.3
SK	71.4	67.2	83.5	74.0
PC	70.6	71.4		71.0

Magnitude Estimation

In this method nine area personnel planners were asked to scale the job areas on need in a direct manner. That is, they were asked to assign numbers to the job areas in accordance with the need for additional men in the job areas. The instructions stressed the fact that the numbers should be chosen in relation to each other. For example, if the need in one job area is just half the need in another job area the numbers assigned to the former should be just half the number assigned to the latter. The scaling questionnaire is presented in Appendix D.

Table 8 describes the data and summary statistics. The utilities in Table 8 have larger average deviations than those obtained through the probability comparison method. However, they also have a greater range. The rank-order correlation of the means is .84.

TABLE 8

NEED RATINGS, MEDIANS, MEANS, AND AVERAGE DEVIATIONS ON TEN JOB AREAS OBTAINED FROM NINE AREA PERSONNEL PLANNERS

Job Areas	Respondents A	B	C	D	E	F	G	H	I	Median Utility	Mean Utility	Average Deviation
SO	100	90	100	10	90	95	100	100	100	100	97.2	3.7
ET	90	100	100	95	100	100	95	90	90	95	95.6	4.0
RM	85	80	95	90	75	80	90	90	80	90	85.0	5.6
YN	70	70	75	50	60	70	55	70	65	70	65.0	6.7
SK	73	50	50	60	55	40	60	65	60	60	57.0	6.0
HM	50	30	80	55	70	50	60	70	50	55	57.2	11.8
MM	75	20	50	60	40	50	65	60	45	50	51.7	11.8
DK	65	10	50	30	20	40	30	60	40	40	38.3	12.4
EN	20	40	50	30	25	30	40	30	55	30	35.5	7.0
PC	10	60	50	20	30	50	40	20	30	30	34.4	13.8

<u>Conversion of the Graduate-Fail Criterion to a Utility Scale</u>

The utility of a graduate was set at the utility of a satisfactory assignee to the corresponding job area--the job areas for which the school trains recruits. Then, the utility of a "fail" was determined in relation to this value. Since the majority of "fails" are sent to the fleet for on-the-job training in the corresponding job, and since those who conduct this training also supervise graduates, the value of the average "fail" was determined by asking the Navy personnel conducting this on-the-job training. The scaling method was magnitude estimation. That is, the supervisors were asked to assume that the average graduate is worth $10,000 to the Navy and to indicate the relative value of a failure. This was done for each job area. The scaling questionnaire is presented in Appendix B.

In this way the utility of a "fail" relative to the utility of a graduate was determined for each school. This scale constitutes a new criterion against which to validate selection tests through the utility function method. Table 9 presents the results for the schools used in the research presented in Chapter VII. The means of the supervisors' responses are in column two, and the number of supervisors in column three. The average deviation is reported rather than the standard deviation because there are extreme deviations, which when squared, would bias estimation of the standard deviation. U_F was set so that U_F is to U_G as Mean Utility of a Fail is to 10,000. The schools are described in Chapter VII.

TABLE 9

THE UTILITY OF A FAIL RELATIVE
TO THE UTILITY OF A GRADUATE

School	Mean Utility of a Fail	N	Average Deviation	U_G*	U_F
SO	$4,864	11	3,008	100	49
ET	4,759	29	1,699	95	45
RM	5,809	54	1,814	90	52
YN	7,316	60	2,558	70	51
SK	6,380	24	2,214	60	38
MM	6,941	76	3,045	50	35
EN	5,886	22	2,502	30	14

*These are the median utilities presented in Table 8.

CHAPTER VI

PAYOFF MATRICES AND A WAY TO DETERMINE THEM

Previous chapters have shown that a new selection test evaluation approach is needed and that statistical decision theory provides a promising model for this problem. However, a formidable prerequisite to using this approach is determining the utilities in payoff matrices. The present chapter is devoted to this task. A way to reduce it to a more manageable form is explained. The payoff matrices used in the next chapter are also presented.

A payoff matrix is a rectangular array of numbers which represent the utilities of decision-outcome combinations. The utilities express the gain and/or loss to the institution in terms of which the decision was made. Thus, they express the desirability of the consequences of decisions. The number may be positive or negative for any particular decision-outcome combination. If it is positive the gain outweighs the loss; while if it is negative the contrary is true.

A payoff matrix for the selection situation is a 2 X 2 matrix of numbers which represent the relative utilities of the four decision-outcome combinations. See for example Figure 7. Thus, for each contingency table with observations in each cell, a payoff matrix is needed with a utility in each cell. A particular utility pertains to each observation in the corresponding cell of the contingency table, indicating its net desirability.

		Decision: Reject	Decision: Accept
Outcome	Satisfactory	U_1	U_2
	Unsatisfactory	U_3	U_4

Fig. 7.--The standard payoff matrix for a dichotomous decision and a dichotomous outcome.

Utility of a Correct Acceptance

In the personnel selection situation the most meaningful element of the payoff matrix seems to be U_2, the utility of a correct acceptance. It can be taken as the utility of obtaining a satisfactory person for the job or position. In many settings this utility might be expressed in dollars through cost accounting or job evaluation procedures.

In other settings construction of a numerical scale by scaling job areas on utility is more efficient. This is particularly true in a large institution where many job areas are to be evaluated and many predictors are used. Even though the resulting utility scale will be "unfamiliar" as compared to the dollar scale, it will have relative meaning across job areas and, therefore, tests. Many psychological scaling techniques are potentially useful for this purpose. Chapter V presents two used in this study.

Utility of an Erroneous Rejection

Rejecting a person who would have been satisfactory may or may not be a serious error. Whether it is or not largely depends on three factors: (1) the need for a satisfactory assignee, (2) the proportion of the testee population that would be satisfactory, and (3) the proportion of the testee population that is needed. U_1 should therefore be some function of U_2, p (the a priori probability), and the proportion needed. The following rules were adopted for the situation to which this study pertains:

1. a satisfactory assignee is lost. $(-U_2)$
2. the loss does not quite balance actually obtaining a satisfactory assignee. $(-U_2 + a)$
3. the loss decreases as p increases.
4. the loss increases as the proportion needed increases (not to be confused with q, the selection ratio).

These rules provide some restriction on U_1. They constitute the leverage the writer was able to bring to bear on this utility in the situation under study. The first rule is true because each time this error of decision occurs a rejected person who would have been satisfactory is in fact lost as far as the assignment is concerned. Rule two modifies rule one. It was adopted because the cost of training an acceptee is not expended on a rejectee. Therefore, the institution loses a satisfactory assignee but saves on training costs as a result of each decision to reject a person who would have succeeded.

Rules three and four were also adopted on logical grounds. Rule three is based on the assumption that the more abundant the persons who would succeed, the less the loss of failing to identify them. At the other extreme, if the test is used to identify rare persons, missing one would be considered a costly error, generally speaking. As with rules one and two, rules three and four are interdependent. Rule four says that the relationship hypothesized in rule three is dependent upon the proportion needed. For example, if very few would succeed, the loss of rejecting such a person might not be great if even fewer are needed.

Following these rules the following was adopted as an arbitrary but reasonable expression of U_1:

$$U_1 = p(1 - q')U_2 - U_2 \qquad (1)$$

or

$$U_1 = -U_2\left[1 - p(1 - q')\right] \qquad (2)$$

The quantity q' is the proportion needed referred to in rule four above. It is not necessarily equal to q, the selection ratio, since the latter is determined by x_c which depends upon the values in the payoff matrix. In the Navy situation q' is the proportion of inductees needed to meet and maintain personnel requirements in the job area. Compared to q it is quite small. The disparity is due to quota restrictions and to the fact that many testees choose another school or fleet duty.

The expression $1 - p(1 - q')$ in Equation (2) is always less than one and represents the absolute size of U_1 relative to U_2. A few examples will show how this expression relates to p and q':

p:	.2	.2	.2	.2	.5	.5	.5	.5	.8	.8	.8	.8
q':	.1	.2	.5	.8	.1	.2	.5	.8	.1	.2	.5	.8
1 - p(1 - q'):	.82	.84	.90	.96	.55	.60	.75	.90	.28	.36	.60	.84

Utility of an Erroneous Acceptance and the Utility of a Correct Rejection

The two remaining quantities of the payoff matrix, U_3 and U_4, are more nebulous in most situations than the two treated above. However, there is a way to circumvent direct estimation.

It was pointed out in Chapter IV that the index U_T is independent of the addition of a constant to the quantities in a row of the payoff matrix. Appendix A presents the mathematical proof. It was also shown in Chapter III that the optimum cutoff is the point on the test score scale where

$$\lambda(x) = \frac{P(F)(U_{R,F} - U_{A,F})}{P(S)(U_{A,S} - U_{R,S})}, \tag{3}$$

assuming that $\lambda(x)$ is a monotonic increasing function of test score, x, and that $U_{R,F} > U_{A,F}$ and $U_{A,S} > U_{R,S}$. Equation (3) was demonstrated in Chapter IV (Equations 9, 10 and 11) to be equivalent to

$$\frac{P(S|x_i)}{P(F|x_i)} = \frac{U_3 - U_4}{U_2 - U_1}. \tag{4}$$

Thus, the optimum cutoff is the point on the test score scale where

$$\frac{P(S)}{1 - P(S)} = \frac{(U_3 - U_4)}{(U_2 - U_1)} . \tag{5}$$

This point may be called the "saddlepoint," a term used in game theory to denote the conditions under which the players win equal amounts on the average. In terms of the accept-reject decision, it is the point on the test score scale where it makes no difference whether the testee is accepted or rejected--the expected payoff is the same in either case.

It is readily apparent from Equation (5) that the difference $U_3 - U_4$ could be computed if an estimate of P(S) for this saddlepoint was available, after the remaining elements, U_1 and U_2, have been estimated. When U_3 is subtracted from the quantities in the bottom row of Figure 7 the matrix in Figure 8 results.

		Reject	Accept
Outcome	Succeed	U_1	U_2
	Fail	0	$U_4 - U_3$
		Decision	

Fig.8.--A modified payoff matrix obtained by adding $-U_3$ to the entries in the bottom row of Figure 7.

Thus, the quantity needed for the calculation of U_T is simply the negative of the difference $U_3 - U_4$ in Equation (5).

If the test has been used for a long time by the institution, in a fairly stable situation, and the cutoff has been set on a trialand-error basis with plenty of feedback from performance criteria, then the established cutoff can be accepted as a fairly accurate estimate of the optimum one. This is the case in the situation to which this study pertains. Therefore, the present cutoff was used, the probability of a graduate at the cutoff determined, and that probability value used to compute $U_3 - U_4$. If the above mentioned conditions were not the case, it would be necessary to obtain an estimate of the saddlepoint. One appealing way would be to ask persons in responsible positions, i.e., ones capable of making higher-level decisions, questions like the following: "Would you assign person P with characteristics m, n, and o, to position A, if he has a 60% chance of success; 70%; 80%; 90%?" This would seem to be a very meaningful task for a person familiar with the current success-ratio and needs of the institution.

Determining the Payoff Matrices for this Study

Table 10 presents the utilities in the payoff matrices for the schools under study. (One school, PC, was dropped because of insufficient data.) Statistics used in determining the utilities are also given: N, the number of students in the sample upon which the other quantities are based; r, the validity coefficient of the selection test, obtained prior to dichotomization and corrected for restriction of the range; p; $P(G|x_c)$, the probability of a graduate at the cutoff; and q'.

Correction of r for restriction of the range was necessary because school assignment had been made on the basis of one or more of the tests to be evaluated, resulting in direct restriction of test score range. A method of correcting for restriction developed by Lawley (1943) and expanded by Meredith (1958), was applied to the data. Data from an unrestricted sample of 500 recruits were used to obtain the base values of the test intercorrelations, means and standard deviations needed for correcting the school matrices. The tests and schools are described in the following chapter.

Each corrected *r* was calculated by computer using the following formula:

$$r = \frac{S_{yx} S_{xx}^{-1} d_x^{\frac{1}{2}} r_{xx}}{\sqrt{S_y^2 + S_{yx} S_{xx}^{-1} d_x^{\frac{1}{2}} (r_{xx} - d_x^{-\frac{1}{2}} S_{xx} d_x^{-\frac{1}{2}}) d_x^{\frac{1}{2}} S_{xx}^{-1} S_{yx}}}; \qquad (6)$$

where

$d_x^{\frac{1}{2}}$ = a vector of the test standard deviations based on an unrestricted sample,

r_{xx} = an intercorrelation matrix of the tests based on an unrestricted sample,

S_{xx} = the variance-covariance matrix of the tests for the restricted sample,

S_{yx} = the vector of validity coefficients for the restricted sample,

S_y^2 = the criterion variance for the restricted sample.

The denominator of Equation (6) is an estimate of the population standard deviation of the criterion. An estimate of the population mean is

$$M = S_{yx} S_{xx}^{-1} (M_x - \bar{x})$$

where S_{yx} and S_{xx} are as defined above, and where

M_x = a vector of test means based on an unrestricted sample,

$\bar{x}$ = the vector of test means for the restricted sample.

These estimates of the population means and standard deviations of the criteria were used in determining the p's in Table 10. The graduate-fail division point is 65 on the school grade scale. The z-score value of this point was calculated for each school by dividing the difference between 65 and the mean by the standard deviation. A table of the normal distribution revealed the proportion, p, above the z-score value.

$P(G|x_c)$ was calculated using "success-ratio" theory as presented by Walker (1957). The success ratio is the probability that persons with a given test score will succeed. Thus, it is the number successful with a given test score, divided by the total number with that test score. Walker developed a "success function" based on the assumptions of normality and linearity that can be used to determine the theoretical success ratio at any particular test score. The success function is the probability function with mean at k/r, k being the z-score value of p on the criterion, and standard deviation equal to $\sqrt{(1 - r^2)}/r$. These are converted to units of the test score scale so that any test score of interest can be located on the success function. (The new mean corresponds to the test mean plus the product of k/r and the test standard deviation. The new standard deviation is equal to the product of $\sqrt{(1 - r^2)}/r$ and the standard deviation of the test.)

TABLE 10

PAYOFF MATRIX UTILITIES AND ANTECEDENT STATISTICS FOR NINE SCHOOL SAMPLES

School	N	r	p	$P(G\|x_c)$	q'	U_2	U_1	$U_4 - U_3$
SO	387	.61	.85	.96	.02	100	-17	-2808
ET	446	.61	.56	.62	.06	95	-45	-228
RM	519	.55	.88	.90	.07	90	-14	-846
YN	417	.45	.97	.99	.05	70	-6	-7524
SK	272	.57	.87	.93	.03	60	-10	-930
HM	403	.62	.99	1.00	.03	55	-2	∞*
MM	554	.68	.90	.97	.07	50	-8	-1875
DK	177	.52	.90	.95	.01	40	-4	-836
EN	214	.63	.80	.90	.04	30	-7	-333

*Due to $P(G|x_c)$ rounding to 1.00

To determine the success ratio at a particular test score, the score is located on the success function, its z-score is computed, and the success ratio is read from a table of the normal distribution.

The success function for each school was determined using the corrected r and the population means and standard deviations for the test and criterion. Then the operational cutoff on the test was located on the success function. $P(G|x_c)$ was obtained from a table of the normal distribution using the z-score value of the cutoff on the success function. See Appendix E.

The proportion of recruits needed, q', was obtained from an unpublished report by the U. S. Navy Personnel Research Activity at San Diego (1964). It is based on the number of men who must enter the job area at the lowest level each year in order for the job area to contain the required number of Petty Officers in the future. This number was then divided by the number of inductees during fiscal 1964.

The U_2's in Table 10 are those obtained through the magnitude estimation scaling method. They are the median utilities presented i The U_1's were calculated using Equation (1) while Equation (5) was used for calculating the $U_4 - U_3$ quantities. In the case of HM, $U_4 - U_3$ is infinity because p is so high and the cutoff is so low that, theoretically, no fails are to be expected.

As can be seen in Table 10, the $U_4 - U_3$ quantities are very large relative to U_1 and U_2. This is due to the fact that $P(G|x_c)$ is quite large for most of the schools. It is not due to the size of U_1, the other quantity upon which $U_4 - U_3$ is based. For example, if U_1 was zero in the case of SO, $U_3 - U_4$ would still be very large, namely 2400. The assumption that the present cutoffs are optimum leads through Equation (5) to the very large $U_4 - U_3$ quantities. To put it another way, use of the present cutoffs implies that the avoidance of failing students is of primary importance to the Navy.

CHAPTER VII

AN EMPIRICAL TRYOUT

The results of an empirical tryout of the new evaluation index U_T are presented in this chapter. The indices $\underline{r}$, $\underline{E}$, and G_U are also presented for comparison. The tests evaluated are the ones used in selecting Navy recruits for technical training. The criterion is final course grade in the correlational analysis; while in the case of U_T it is dichotomized final grade, graduate-fail. Final grade is assigned by school instructors through a differential weighting of the individual achievement and proficiency tests taken during the course of training.

The Schools Sampled

1. Sonarman (SO).--This is a 16-week course. The curriculum consists of (1) operation of sonar equipment, (2) International Morse Code communications, (3) basic electricity, electronics and sonar equipment circuitry, (4) cleaning and lubrication of sonar equipment, and (5) use of equipment for testing electronic performance of sonar equipment.

The final grade is based on four practical and 17 written examinations. The latter receive 80 per cent of the weight. (This is the formal weighting system. The effective weights which depend on variances and intercorrelations are not known.)

The a priori probability (the proportion who would graduate if selection was random) has been estimated in previous research to be .85. The cutoff used for selecting recruits for this school is GCT + ARI = 110 which is .52 standard deviations above the mean. (See below for a description of these tests.)

2. Electronics Technician (ET).--This school is 38 weeks long. The curriculum covers basic electricity and electronics, required mathematics, and maintenance and repair of communication equipment.

The final grade is based on eight practical and 18 written examinations. The latter receive 88 per cent of the weight.

The a priori probability, *p*, for this school is .56. The present cutoff is GCT + ARI - ETST = 170 which is .72 standard deviations above the mean.

3. Radioman (RM).--This is a 24-week school. Its curriculum consists of instruction in the operation of radios, teletypewriters and voice radio equipment, transmission and reception of messages by International Morse Code, basic electricity and electronics, operation and maintenance of receiving and transmitting equipment.

Final grade is based on 39 examinations. Approximately one-half of these are written and one-half are practical. Written and practical examinations are weighted equally in arriving at final grade.

For this school *p* is .88. The cutoff is GCT + ARI = 100 which is .15 standard deviations below the mean.

4. Yeomen (YN).--This school is eight weeks long. The curriculum covers clerical duties, including typing, filing, operation of duplicating machine equipment and general office work, and records for courts-martial.

Written examinations (7) cover 80 per cent of the final grade and practical examinations (6) make up the remaining 20 per cent.

For this school p is .97. The cutoff is GCT + CLER = 110 which is .97 standard deviations above the mean.

5. Storekeepers (SK).--This is a 12-week course. The instruction covers general stores supply afloat, clothing and small stores, ships store, provision, repair parts, records and reports, typing, and practical work in all phases of supply afloat.

The final grade is based on one practical and 21 written examinations. The latter receive 88 per cent of the weight.

The a priori probability, p, is .86, and the cutoff is GCT + ARI = 105 which is .18 standard deviations above the mean.

6. Machinist's Mates (MM).--This school consists of 12 weeks of instruction in principles of main propulsion machinery and auxiliaries operation, maintenance and repair; handtools, gauges and instruments as used in operating, checking, adjusting and performing preventive maintenance. Auxiliary machinery covered includes refrigeration equipment, evaporators, pumps, compressors, heat exchangers, and emergency electrical generators.

The final grade is based on 12 practical and 61 written examinations. The latter receive a formal weight of 93 per cent.

For this school p is .90. The cutoff is GCT + MECH = 105 which is .29 standard deviations above the mean.

7. Engineman (EN).--This is a 12-week school for the training of men to operate, maintain, and repair internal-combustion engines. The school provides for study and work experience in the following areas: (1) Mathematics, blueprint reading, temperature and pressure instruments, and basic electricity; (2) Threadcutting, pipefitting, soldering and use of hand tools; (3) Theory, construction, and operation of diesel and gasoline engines and their associated equipment; (4) Auxiliaries including boilers, distilling plants, air compressors, pumps, refrigeration, and air conditioning; (6) Damage control.

The final grade is based on application marks and 61 written examinations, the latter receiving 95 per cent of the weight.

For this school p is .80. The cutoff is ARI + MECH = 105 which is .37 standard deviations above the mean.

The Selection Tests

1. The General Classification Test (GCT).--This is a 100-item test of verbal aptitude consisting of sentence completion and verbal analogy items. The alternate form reliability is .93. A single Navy Standard Score, having a mean of 50 and a standard deviation of 10, was used.

2. The Arithmetic Test (ARI).--This test consists of two separately-timed subtests. A 20-item Arithmetic Computation subtest

provides a measure of speed and accuracy in performing elementary computations, and a 30-item Arithmetic Reasoning subtest provides a measure of ability to solve verbally presented quantitative problems. Only total score in Navy Standard Score was used. The alternate form reliability of this test is .85.

3. The Mechanical Test (MECH).--This test consists of two separately-timed 50-item subtests: Mechanical Comprehension and Tool Knowledge. Only total score was used; and it was expressed in Navy Standard Score. The alternate form reliability is .86.

4. The Clerical Test (CLER).--This is a 210-item highly speeded test of number matching. The subject compares two adjacent columns of 5-to 9-digit numbers and indicates whether or not they are identical. A total score in Navy Standard Score form was obtained using the formula Number Right minus Number Wrong. The alternate form reliability of this test is .77.

5. The Electronics Technician Selection Test (ETST).--This test is primarily a measure of achievement and experience in areas related to electronics maintenance. It has five separately-scored subtests: Mathematics (20 items, some requiring a knowledge of algebra for their solution); Science (20 items, primarily physics); Shop Practice (10 items); Electricity (15 items) and Radio (15 items, some requiring a knowledge of electronic circuitry). Total test score in Navy Standard Score was used. The alternate form reliability of this test is .89.

The Results

The results are presented in Tables 11 and 12. In Table 11 the three evaluation indices are presented along with pertinent information on the samples and schools. The a priori probabilities are those in Table 10 in Chapter VI. The selection ratio, q, was obtained in each case from a table of the normal distribution using the σ-value of the cutoff given above. N is the total sample size upon which the calculations involved in the correlational approach are based. Only n_2 and n_4 were used in calculating G_U and U_T. G_U was obtained using Equations (13) and (15) in Chapter IV, the utilities in Table 9, and the p and n's in Table 11. U_T was obtained using Equations (4) and (7) in Chapter IV, the utilities in Table 10, and the p and n's in Table 11. The discrepancies between N and $n_2 + n_4$ are due to waivers, i.e., acceptees whose test score composite did not exceed the cutoff.

G_U and U_T are expressed in units of the same utility scale, the one obtained through the scaling techniques described in Chapter V. However, G_U expresses the utiles gained for each man selected by the tests while U_T expresses the utility of the tests for all the accept-reject decisions which led to this sample of students above the cutting score.

$G_{U_{Max}}$ and $U_{T_{Max}}$ are maximum values, i.e., ones which would be obtained with perfect selection (if all the selectees had graduated). They are the utility function and decision-theoretic approaches' counterparts of an r of 1.00. When $p \geq q$ the formulas are

$$G_{U_{Max}} = U_G - pU_G + (1 - p)U_F$$

TABLE 11

THE THREE EVALUATION INDICES AND CLOSELY RELATED STATISTICS FOR SAMPLES FROM SEVEN NAVY SCHOOLS

School	N	p	q	n_2	n_4	Evaluation Index			$G_{U_{Max}}$	$U_{T_{Max}}$	P_G above x_c	P(G) above x_c
						r	G_U	U_T				
SO	387	.85	.30	289	30	.61	2.85	52212	7.65	139960	.906	.98
ET	446	.56	.24	164	19	.63	16.81	22639	22.00	29631	.896	.86
RM	519	.88	.56	334	40	.55	0.50	4587	4.56	42187	.893	.97
YN	417	.97	.17	236	4	.45	0.25	24320	0.57	54720	.983	.99
SK	272	.86	.43	169	18	.57	0.96	8180	3.08	26180	.904	.97
MM	554	.90	.38	318	4	.68	1.31	54511	1.50	62243	.988	.99
EN	214	.80	.35	159	8	.63	2.43	9398	3.20	12358	.952	.97

Notes:

N: sample size.
p: a priori probability (or base rate).
q: selection ratio.
n_2: number graduating above the cutoff x_c.
n_4: number failing above the cutoff x_c.
r: product-moment correlation coefficient obtained prior to dichotomization and corrected for restriction of range using the Lawley method described in the preceding chapter.
G_U: test evaluation index of the utility function approach.
U_T: test evaluation index of the decision-theoretic approach.
$G_{U_{Max}}$: the value of G_U that would have been obtained had the test provided perfect selection.
$U_{T_{Max}}$: the value of U_T that would have been obtained had the test provided perfect selection.
P_G above x_c: the proportion graduating above the cutoff x_c.
P(G) above x_c: The probability of a graduate above x_c obtained from the Taylor-Russell tables, or $P(G|x > x_c)$.

TABLE 12

THE UTILITY OF SELECTION TESTS AS ESTIMATED BY THE THREE APPROACHES: CORRELATIONAL, UTILITY FUNCTION, AND DECISION-THEORETIC

School	Proportion Improvement Over Chance Prediction			The Index Values Expressed In Number of Graduates		
	E	$G_U/G_{U_{Max}}$	$U_T/U_{T_{Max}}$	$[P(G)^* - p](n_2 + n_4)$	$G_U(n_2 + n_4)/U_G$	U_T/U_G
SO	.21	.37	.37	41.5	9.1	522.1
ET	.23	.76	.76	54.9	32.4	238.[illegible]
RM	.16	.11	.11	33.7	2.1	51.[illegible]
YN	.11	.44	.44	4.8	0.9	347.4
SK	.18	.31	.31	20.6	3.0	136.[illegible]
MM	.27	.88	.88	28.3	8.5	1090.[illegible]
EN	.23	.76	.76	28.4	13.6	313.[illegible]

$^*P(G|x > x_c)$

and

$$U_{T_{Max}} = (n_2 + n_4)(U_2 - U_1) - U_c.$$

These were derived from the formulas for G_U and U_T by making the following substitutions in Equations (13) and (4) in Chapter IV: $n_2 = n_2 + n_4$ and $n_4 = 0$. When $q > p$ the substitutions would be $n_2 = pN$ and $n_4 = N(q-p)$ yielding different formulas for $G_{U_{Max}}$ and $U_{T_{Max}}$.

The last two columns of Table 11 present the obtained and theoretical proportions graduating above the cutoff. P_G above x_c is $n_2/(n_2 + n_4)$ while P(G) above x_c is the probability that a randomly selected person above the cutting score will graduate. The latter was obtained from the Taylor-Russell tables using r, p, and q for each school.

Taking SO as an example, the validity coefficient is .61, G_U is 2.85, the gain in utiles for each man selected by the test, and U_T is 52,212, the utility of the test for the accept-reject decisions which led to this sample of students.

In Table 12 the results are presented in two forms which make the three evaluation indices comparable. They pertain to utility or practical significance. In terms of the proportion improvement over chance prediction the utility function approach and the decision-theoretic approach agree precisely. $U_T/U_{T_{Max}}$ and $G_U/G_{U_{Max}}$ are larger than E in all but one of the seven schools.

In the last three columns of Table 12 the three evaluation indices are expressed in terms of "number of graduates." Only in the first of these columns, which pertains to the correlational approach, can the quantities be taken as the number of graduates actually gained through the use of the test. This is the Taylor-Russell interpretation of validity coefficients. In the other two columns the utility index values have been translated into "number of graduates" by dividing them by U_G, the utility of a graduate of that school. For example, the first term of the last column was obtained by dividing U_T for SO--52,212--by 100, the utility of a graduate of SO school.

Thus each quantity in these three columns means that the selection test was as valuable as would be actually adding that many graduates to the operational Navy. Any differences between the three quantities in a given row must be due to differences in the evaluation approaches which lie behind the three test evaluation indices. As can be seen there are large differences. The approaches lead to radically different conclusions regarding the utility of the tests. The decision-theoretic approach demonstrates that the tests are worth much more than either the correlational approach or the Taylor-Russell approach would indicate. In the case of SO, the Taylor-Russell approach indicates the use of the selection tests meant a gain of 41.5 SO graduates for this group of 319--$n_2 + n_4$--selectees. The utility function approach indicates the gain was equivalent to gaining 9.1 new SO graduates. The decision-theoretic approach indicates the gain was equivalent to gaining 522.1 new SO graduates.

Some of the quantities in the last column of Table 12 are quite large, indicating the tests were worth much more than would be expected. This is primarily due to the fact that $U_4 - U_3$ is very large for these schools. (See Table 10 in Chapter VI.) Any test that reduces the number of erroneous acceptees in such a situation will have high utility. In the case of MM the numerical reduction was 28. As was pointed out in the concluding paragraph of Chapter VI, the enormity of the $U_4 - U_3$ quantities is due to the fact that $P(G|x_c)$ is very close to 1.00 for all but one of the schools: ET. Since these $U_4 - U_3$ quantities follow mathematically from the model presented in Chapter III and the assumption that the present cutoffs are optimum, if the model is appropriate for test-based selection decisions these results make the cutoffs suspect.

Since the utilities and marginal probabilities are peculiar to the empirical situation, care should be exercised in generalizing these results.

CHAPTER VIII

DISCUSSION

In this chapter six questions which are pertinent to this study are raised and discussed. They are:

1. Should the complete payoff matrix for selection contain zeros in the reject column?

2. What changes does a valid test make in the "chance" 2 X 2 table and how does U_T relate to them?

3. How do the payoff matrices needed for meaningful test evaluation relate to those needed in decision making?

4. How does practical significance differ from statistical significance?

5. How valuable is U_T for comparing tests?

6. Should the $\underline{U}$ scale be converted to a dollar scale?

These questions are discussed in order.

1. Should the complete payoff matrix for selection contain zeros in the reject column?

Cronbach and Gleser (1957) make no distinction between rejectees as far as utility is concerned, apparently assuming that since a rejectee is not in the institution the decision to reject him can have no positive or negative effect on the institution. However, the evaluation of a test should involve consideration of the consequences of errors and correct decisions. If persons who know the needs of the institution decide that rejecting a person who would have done well in a particular assignment is a loss and a serious error, then the test should be evaluated in terms of erroneous rejections; and values should appear in that cell of the payoff matrix. In other words, the answer to this quest' n must depend on the empirical situation. There is no substitute for intra-institutional analysis of the gains and losses resulting from decisions.

There are however ways to make these analyses easier and to give them stability. Chapter III presents a system in terms of the optimal cutoff, the one which maximizes payoff. Prior to developing this rationale and these mathematical relationships, the author of this dissertation tried many logical analyses and found himself developing many diverse payoff matrices which would have led to quite different conclusions as to the utility of selection tests. Thus, even within an institution the determination of the most appropriate payoff matrix is very difficult without some rationale and system to guide the analysis.

The system presented in Chapter III is based on decision theory. This theory assumes that decisions should be made in such a way as to maximize payoff. When faced with a choice, the assumption is that the best course of action is that which will, on the average, lead to the greatest payoff. The decision to "accept" is made only when the expected payoff is greater than the expected payoff for "reject." The "saddlepoint" is the optimal cutoff, the point on the test score scale where the expected payoffs are equal. This theory clearly assumes nonzero quantities in the reject column. The "saddlepoint" and maximum payoff concepts are meaningless otherwise. While the answer to the question heading this section should be empirically determined in every case, the logic of decision theory strongly supports the use of complete payoff matrices.

2. What changes does a valid test make in the "chance" 2 X 2 table and how does U_T relate to them?

If selection was random a "chance" 2 X 2 table would result. In terms of a table of this nature the function of a valid test is figuratively to shift persons from one cell to another. In Table 13 the arrows represent these shifts.

TABLE 13

A "CHANCE" 2 X 2 TABLE SHOWING THE FIGURATIVE SHIFTS OF PERSONS A VALID TEST WOULD MAKE

Outcome		Decision: Reject		Decision: Accept
	Graduate	28	→	12
	Fail	42	←	18

The payoff matrix indicates just what these shifts are worth to the institution. Consider the payoff matrix presented in Table 14. A shift of a person in the top row of the corresponding 2 X 2 table is worth 18 units to the institution, since instead of suffering a 10-

unit loss it gains 8 units. Similarly, a shift in the bottom row is worth 11 units.

If the obtained 2 X 2 table presented in Table 15 is compared with the "chance" table in Table 13, the figurative shifts are 10 in the top row and 10 in the bottom row. These are worth 180 + 110 = 290 units to the institution. This is also the value of U_T. If the complete "chance" and obtained 2 X 2 tables are available, this procedure can be used instead of the formula for U_T given in Chapter IV. The result would be the same in every case.

TABLE 14

A HYPOTHETICAL PAYOFF MATRIX

		Decision: Reject	Decision: Accept
Outcome	Graduate	-8	10
	Fail	6	-5

TABLE 15

A HYPOTHETICAL CONTINGENCY TABLE

		Decision: Reject	Decision: Accept
Outcome	Graduate	18	22
	Fail	52	8

3. <u>How do the payoff matrices needed for meaningful test evaluation relate to those needed for decision making?</u>

The two previous sections make obvious the importance of the absolute size of the values in the payoff matrices for test evaluation. And, as is pointed out in the next section, these values must have meaning within a particular institutional setting. To put it succinctly, determining the utility of a test for a particular decision requires a quantitative estimate of the gain to the institution using it. This estimate is directly affected by the absolute size of the values in the payoff matrix.

If on the other hand, the payoff matrix is to be used only for decision-making--and in testing this means determining the optimum cutoff--the relative size of the values in the payoff matrix is all that is needed. Proportionally equal reductions or increases will not affect the outcome.

The values in the payoff matrices in this study were given quantitative and institutional meaning by scaling the job areas on need--by determining the relative need for more men in the job areas. These values appear in the "correct acceptance" cell of the payoff matrices. The other values in the payoff matrices were determined in relation to these values. The quantitative estimate of utility, namely U_T, is expressed in the units of this need scale.

4. <u>How does practical significance differ from statistical significance?</u>

Practical significance refers to value or utility while statistical significance refers to the probable stability of an obtained statistic. In selection the statistical significance of a correlation coefficient indicates that there is probably a reliable association between the selection test and the criterion. The fact of statistical significance has no implication of how much association there is between these factors. It simply means there is probably <u>some</u> association.

Practical significance, on the other hand, corresponds more closely to the common man's concept of significance, namely, important and valuable. Instead of referring to an abstract level of confidence, e.g., 95%, it refers to concrete utility. This means that practical significance refers to the utility of the selection test in a particular situation as well as for a particular decision. All the situational (institutional) factors which are affected by the consequences of the decision should be reflected. This contrasts with statistical significance of validity coefficients which is divorced from the situation except as it relates to the criterion. The end result is an estimate of the test's utility to the institution employing it as contrasted with an estimate of the criterion variance accounted for, or of predictive efficiency.

5. <u>How valuable is U_T for comparing tests?</u>

As was pointed out in Chapter II, the primary statistic being used to evaluate selection tests is the correlation coefficient. Of two tests with equal selection ratios, the test which correlates highest with a dichotomous criterion will be best by any standard (disregarding cost of testing). This is because there is but one degree of freedom in a 2 X 2 table with fixed marginals. A decrease in the number of a particular type of error must be accompanied by an equal decrease in the other type of error as well as an equal increase in the other two cells. Thus, with any payoff matrix a decrease in the number of errors, regardless of which one, will increase the correlation coefficient as well as U_T. With fixed marginals, any change in the 2 X 2 table will affect both statistics in the same way. In any particular selection situation, such as selection for one of the Navy schools, they would lead to the same choice of test. Although this was not investigated in this study, it is likely that U_T would be superior for two reasons: (1) it would provide a quantitative estimate of how much better one test is than another in a particular situation and (2) this quantitative estimate would be in terms of a utility scale having broad meaning for a particular institution and to which the above aspects can be related. Correlation coefficients on the other hand cannot claim to indicate how much better one test is than another for a particular decision and institution because the needs and costs, gains and losses, peculiar to that institution are not reflected in them.

6. Should the U scale be converted to a dollar scale?

Two general approaches to utility analysis seem obvious. The one taken in this study is to establish a purely general utility scale. The other is to use the familiar dollar scale. The former approach was used because the author imagined that it would be very difficult for the respondents of the questionnaires to place the Navy's needs on the dollar scale. Intra-individual conflict and hightened inter-individual variation seemed likely.

After the scale values on the general scale have been obtained it would probably be quite easy in most institutional settings to make accurate links between the utility scale and the dollar scale. This might be done through cost-accounting or judgemental procedures. It would greatly increase the meaningfulness of the scale and put many intra- and inter-institutional relationships on a quantitative basis.

CHAPTER IX

SUMMARY AND CONCLUSIONS

The correlational approach to selection test evaluation was examined and found to have serious limitations. An approach based on statistical decision theory was developed. Two new methods were presented, one called the utility function method and the other the decision-theoretic method. The former is largely based on Brogden's work and involves the comparison of criterion groups in terms of their utility to the institution using the selection test being evaluated. The decision-theoretic method is based on statistical decision theory and involves the construction of a payoff matrix corresponding to the contingency table relating the test to the criterion. The cell frequencies are weighted in a utility equation by the payoff values in the corresponding cells of the payoff matrix. This utility equation represents a new test evaluation index which directly expresses the utility of the test to the institution using it.

Both of these new methods require the measurement of values peculiar to the institution using the test. The utility function method requires that the performance criterion be translated to a utility function; while the decision-theoretic method requires that a payoff matrix be developed which reflects the gains and losses each cell observation represents to the institution.

The three approaches (correlational, utility function, and decision-theoretic) were compared with tests used to select students for technical schools in the U. S. Navy. Scaling techniques were developed for the measurement of values inherent in the Navy situation. Specifically, the graduate-fail criterion was translated to a utility scale and the job areas were scaled on need (or the utility of graduates to the Navy). Using scale values obtained for the job areas, a payoff matrix was constructed for each school on the assumption that the presently used test cutoffs are optimal.

The three approaches led to quite different indications regarding the utility of the selection tests evaluated. The two new methods agreed in terms of the proportion improvement over chance prediction provided by the tests while the correlational approach tended to underestimate this proportion. In terms of practical significance the decision-theoretic approach lead to much more positive conclusions regarding the tests than did the other two approaches.

In addition to the above, perhaps the following conclusions can be drawn from this study:

(1) Statistical decision theory is particularly well suited for the usual test evaluation situation.

(2) Psychological scaling methods provide a solution for the measurement of values required in the application of the decision-theoretic approach to test evaluation.

(3) Supplementation of correlational analysis of tests with decision-theoretic analysis is likely to lead to new insights into the utility of tests for personnel decisions.

BIBLIOGRAPHY

Brogden, H. E. On the interpretation of the correlation coefficient as a measure of the predictive efficiency. J. educ. Psychol., 1946, 2, 171-182.

Brogden, H. E. A new coefficient: Application to biserial correlation and to estimation of selective efficiency. Psychometrika, 1949, 14, 169-182.

Buros, O. K. The fifth mental measurements yearbook. Highland Park, N. J.: Gryphon Press, 1959.

Chernoff, H., & Moses, L. E. Elementary decision theory. New York: John Wiley and Sons, Inc., 1959.

Cronbach, L. J., & Gleser, G. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1957.

Dorcus, R. M., & Jones, M. H. Handbook of employee selection. New York: McGraw-Hill Book Co., Inc., 1950.

Good, H. H. Deferred decision theory. In R. E. Machol & P. Gray (Eds.), Recent developments in information and decision processes. New York: Macmillan, 1962.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill Book Co., Inc., 1956.

Lawley, D. N. A note on Karl Pearson's selection formulae. Proc. roy. Soc. Edenburgh, Sec. A, 1943, 62, Part I, 28-30.

Marschak, J. Probabilities in the social sciences. In P. F. Lazarsfeld (Ed.), Mathematical thinking in the social sciences. Glencoe, Ill.: Free Press, 1954.

Meredith, W. M. The estimation of criterion parameters from a biased sample. A Report Prepared by the University of Washington Division of Counseling and Testing Services. Seattle: University of Washington, 1958.

Swetts, J. A., Tanner, W. P., Jr., & Birdsall, T. G. Decision processes in perception. Psychol. Rev., 1961, 68, 301-340.

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection. Journal of Applied Psychol., 1939, 23, 565-578.

U. S. Navy Personnel Research Activity. Striker planning ratios for Navy ratings. San Diego, Calif., 1961.

Walker, D. A. The theory and use of the success-ratio. Brit. J. statist. Psychol., 1957, 10, 105-111.

APPENDIX A. ALTERATION OF THE PAYOFF MATRIX

Theorem: The difference $U - U_c$ is independent of the addition of any constant to the values of both entries in a row of the payoff matrix where

$$U = n_1U_1 + n_2U_2 + n_3U_3 + n_4U_4, \tag{1}$$

$$U_c = (p - pq)NU_1 + pqNU_2 + (1 - p - q + pq)NU_3 + (q - pq)NU_4, \tag{2}$$

$$N = n_1 + n_2 + n_3 + n_4,$$

and where the contingency matrix is

$$\begin{array}{cc|c} \begin{array}{|c|c|} \hline n_1 & n_2 \\ \hline n_3 & n_4 \\ \hline \end{array} & & \begin{array}{c} p \\ 1 - p \end{array} \\ \begin{array}{cc} 1 - q & q \end{array} & & \end{array}$$

and the payoff matrix is

$$\begin{array}{|c|c|} \hline U_1 & U_2 \\ \hline U_3 & U_4 \\ \hline \end{array}$$

Proof: Consider the following matrix:

$U_1 - k$	$U_2 - k$
U_3	U_4

The *U* for it is

$$U = n_1(U_1 - k) + n_2(U_2 - k) + n_3U_3 + n_4U_4 \tag{3}$$

and U_c is

$$U_c = (p - pq)N(U_1 - k) + pqN(U_2 - k)$$
$$+ (1 - p - q + pq)NU_3 + (q - pq)NU_4. \tag{4}$$

Since the last two terms of Equations (3) and (4) are the same as the corresponding terms of Equations (1) and (2) respectively, proof of the theorem involves showing that

$$n_1U_1 + n_2U_2 - \left[(p - pq)NU_1 + pqNU_2\right] = n_1(U_1 - k) + n_2(U_2 - k)$$
$$- \left[(p - pq)N(U_1 - k) + pqN(U_2 - k)\right].$$

Simplifying

$$n_1U_1 + n_2U_2 - pNU_1 + pqNU_1 - pqNU_2 = n_1U_1 - n_1k + n_2U_2 - n_2k - pNU_1$$
$$+ pNk + pqNU_1 - pqNk - pqNU_2 + pqNk.$$

Canceling yields

$$0 = -n_1k - n_2k + pNk. \tag{5}$$

Now since

$$p = \frac{n_1}{N} + \frac{n_2}{N}$$

Equation (5) can be written

$$0 = -n_1 k - n_2 k + \left(\frac{n_1}{N} + \frac{n_2}{N}\right) NK$$

$$0 = -n_1 k - n_2 k + n_1 k + n_2 k$$

$$0 = 0.$$

APPENDIX B. THE QUESTIONNAIRE USED IN CONVERTING THE GRADUATE-FAIL CRITERION TO A UTILITY SCALE

INTRODUCTION

In today's technically advancing Navy, personnel policies must be kept up to date. You can help in this task by answering the questions which make up this questionnaire. This information will be considered by the Chief of Naval Personnel, along with other information obtained from other sources, in revising personnel policies and practices.

The questions deal with certain of your experiences and opinions regarding Navy training in your own rating. The answers you provide will be used only for research purposes and will in no way affect you as an individual. Please answer all questions even though you can provide only a rough guess on some.

Answer the questions on your own. Do not discuss them with others. Your judgment is important to this research and to the Navy.

A. Identification and Background Information

1. Name______________________________________
   Last　　　First　　　Middle

2. Service Number______________ 3. Pay Grade________________

4. Rating (ET, YN, etc.) ______ 5. Ship or Station______________

   ____

6. Indicate your attendance at schools for your present rating:

School	Attended?	If yes, did you graduate?	
A-School	Yes___ No___	Yes___	No___
B-School	Yes___ No___	Yes___	No___
C-School	Yes___ No___	Yes___	No___

7. How long have you been in your present rating?_______years.

8. How much of the above time were you engaged in the duties of your present rating?________years.

9. Approximately how many men in your rating have you supervised for an extended period, say 3 months or more?________(total number during your career).

10. Approximately how many of those you supervised were graduates of the A-School for your rating?___________.

11. Approximately how many men who were dropped from the A-School for your rating because of failing grades have you worked with, supervised, or trained on the job?__________.

12. Approximately how many strikers in your rating who had no A-School training have you worked with, supervised, or trained?
_____________________________.

13. Have you been an instructor for your rating in A-School______; B-School___________; C-School__________?

B. Judgments Regarding Training in Your Rating

In this section you are to compare graduates and dropouts (failures) from the A-School for your rating. You are asked to judge them in terms of their value to the Navy during their first enlistment. Consider their contribution to the efficiency and capability of the Navy.

1. Assuming that the average graduate of the A-School for your rating is worth $10,000 to the Navy during his first enlistment, how much is the average dropout from that school worth who receives on-the-job training in your rating? (The time period to be considered in both cases is the 4 years of their first enlistment.)

$___________.00

Notice that you are to consider only some dropouts, namely, only those who later receive on-the-job training in your rating. (For the purpose of this questionnaire assume that you personally did not conduct this training.) Try to estimate the average dropout's over-all value to the Navy within your rating, not just his value on a particular piece of equipment or on a subtask within the rating. Use the $10,000 figure as a guide or standard.

2. Now consider those persons in your rating who never had any School training--they went directly to the fleet after recruit training and became strikers in your rating. How much is the average non-school striker in your rating worth to the Navy

during his first enlistment? As before, use the $10,000 figure for graduates as a guide or standard.

$__________.00

APPENDIX C. THE QUESTIONNAIRE USED IN THE PROBABILITY COMPARISON SCALING METHOD

CLASSIFICATION INTERVIEWER OPINION SURVEY

Name____Billet________________
Last First Initial

Rate/Rank________Years in present billet______________

Years Service__________Years experience in classification___________

What this is about

In today's fast changing Navy, personnel policies must be kept up-to-date. You can help in this important task by answering the questions below.

This questionnaire deals with the classification of recruits for assignment to Class "A" schools. Its purpose is to discover what classification decisions you would make in a series of artificial situations. Your responses will be combined with those of other classifiers in an attempt to discover what pattern of decisions are made by a group of experienced classification interviewers.

This questionnaire is being given only for research purposes at the present time. No participant will be identified by name or in any other way in the research reports.

PART I

Directions for Part I

Each question refers to a Class "A" school and to an imaginary recruit who is to be classified.

As everyone knows, you can not be absolutely sure that every recruit you send to a school will do well. You no doubt attempt to determine each recruit's chances of success in various schools during the interview. Assume in each question below that you have decided on the basis of his test scores, interests and experience that he is best suited for the rating to which the school corresponds.

You are to indicate whether you would send him to that school or not if you think he has the chance of success stated in the question;

"success" means graduating without being set back or singled out for an undue amount of tutoring.

Your judgements will probably reflect differences in school quotas and shortages in the ratings. Try to assume a stable quota situation, using as the basis for your judgements the average situation as it existed during 1963.

Sample question

Would you send him to Electronic Technician school if you think he has

	Yes or No
(a) a 50% chance of success?	no
(b) a 60% chance of success?	no
(c) a 70% chance of success?	yes
(d) an 80% chance of success?	yes
(e) a 90% chance of success?	yes

The person who answers this question in this way would not send a recruit to ET school if he believes the recruit has a 50% or a 60% chance of success in that school. He would send him to ET school if he believes the recruit has at least a 70% chance of success in that school.

The questions

Please answer the following questions, writing "yes" or "no" in each blank as was done in the sample question.

1. Would you send him to Electronics Technician school if you think he has

	Yes or No
(a) a 50% chance of success?	______
(b) a 60% " " "	______
(c) a 70% " " "	______
(d) an 80% " " "	______
(e) a 90% " " "	______

2. Would you send him to Storekeeper school if you think he has — Yes or No

(a) a 50% chance of success? ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

3. Would you send him to Radioman school if you think he has — Yes or No

(a) a 50% chance of success? ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

4. Would you send him to Postal Clerk school if you think he has — Yes or No

(a) a 50% chance of success? ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

5. Would you send him to Hospital Corpsman school if you think he has — Yes or No

(a) a 50% chance of success? ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

6. Would you send him to Machinist's Mate school if you think he has — Yes or No

(a) a 50% chance of success? ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

7. Would you send him to Disbursing Clerk school if you think he has — Yes or No

(a) a 50% chance of success? ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

8. Would you send him to Sonarman school if you think he has — Yes or No

(a) a 50% chance of success ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

9. Would you send him to Engineman school if you think he has — Yes or No

(a) a 50% chance of success ______

(b) a 60% " " " ______

(c) a 70% " " " ______

(d) an 80% " " " ______

(e) a 90% " " " ______

10. Would you send him to Yeoman school if you think he has

	Yes or No
(a) a 50% chance of success?	______
(b) a 60% " " "	______
(c) a 70% " " "	______
(d) an 80% " " "	______
(e) a 90% " " "	______

PART II

<u>Directions for Part II</u>

Each question refers to <u>two</u> Class "A" schools and to an imaginary recruit who <u>must</u> be sent to one of the two schools. (Assume that you have decided on the basis of his test scores, interests and experience that he should be sent to one or the other of these schools.)

In the questions below, the recruit's chances of success are stated. You are to indicate your preference of assignment by placing a mark in one of the two spaces <u>in each line</u>. Since your preferences might vary with changes in quotas and with shortages in the ratings, take the average conditions during 1963 as the circumstances for your judgments.

<u>Sample question</u>

To which school would you assign a recruit if you think his chances of success are

	Yeoman			Electronics Technician	
(a)	80%	✓	and	60%	
(b)	80%		and	70%	✓
(c)	80%		and	80%	✓
(d)	80%		and	90%	✓
(e)	80%		and	95%	✓

The person who answers this question in this way believes it would be better to assign a recruit to Yeoman school than to Electronics Technician school if he has an 80% chance of success in YN school and a 60% chance of success in ET school--in other words, if the chances stated in (a) are true. If the recruit's chances of success are 80% for YN school and 70% or above for ET school, this person would prefer to assign the recruit to ET school.

<u>The questions</u>

Please answer the following questions making a check in one of the blanks in each line as was done in the sample question. Take your time, resting frequently if the task seems difficult. As in Part I, "success" means graduating without being set back or singled out for a lot of special tutoring.

1. To which school would you assign a recruit if you think his chances of success are

	Radioman		Engineman
(a)	70% _____	and	60% _____
(b)	70% _____	and	70% _____
(c)	70% _____	and	80% _____
(d)	70% _____	and	90% _____
(e)	70% _____	and	95% _____

(Be sure you made 5 marks--one for each pair of percentages)

2. To which school would you assign a recruit if you think his chances of success are

	Postal Clerk		Yeoman
(a)	90% _____	and	60% _____
(b)	90% _____	and	70% _____
(c)	90% _____	and	80% _____
(d)	90% _____	and	90% _____
(e)	90% _____	and	95% _____

3. To which school would you assign a recruit if you think his chances of success are

	Radioman		Hospital Corpsman
(a)	70% _____	and	60% _____
(b)	70% _____	and	70% _____
(c)	70% _____	and	80% _____
(d)	70% _____	and	90% _____
(e)	70% _____	and	95% _____

4. To which school would you assign a recruit if you think his chances of success are

	Disbursing Clerk		Hospital Corpsman
(a)	80% ______	and	60% ______
(b)	80% ______	and	70% ______
(c)	80% ______	and	80% ______
(d)	80% ______	and	90% ______
(e)	80% ______	and	95% ______

5. To which school would you assign a recruit if you think his chances of success are

	Machinist's Mate		Sonarman
(a)	90% ______	and	60% ______
(b)	90% ______	and	70% ______
(c)	90% ______	and	80% ______
(d)	90% ______	and	90% ______
(e)	90% ______	and	95% ______

6. To which school would you assign a recruit if you think his chances of success are

	Storekeeper		Engineman
(a)	90% ______	and	60% ______
(b)	90% ______	and	70% ______
(c)	90% ______	and	80% ______
(d)	90% ______	and	90% ______
(e)	90% ______	and	95% ______

7. To which school would you assign a recruit if you think his chances of success are

	Engineman		Machinist's Mate
(a)	80% ______	and	60% ______
(b)	80% ______	and	70% ______
(c)	80% ______	and	80% ______
(d)	80% ______	and	90% ______
(e)	80% ______	and	95% ______

8. To which school would you assign a recruit if you think his chances of success are

	Postal Clerk		Engineman
(a)	90% _____	and	60% _____
(b)	90% _____	and	70% _____
(c)	90% _____	and	80% _____
(d)	90% _____	and	90% _____
(e)	90% _____	and	95% _____

9. To which school would you assign a recruit if you think his chances of success are

	Electronics Technician		Radioman
(a)	70% _____	and	60% _____
(b)	70% _____	and	70% _____
(c)	70% _____	and	80% _____
(d)	70% _____	and	90% _____
(e)	70% _____	and	95% _____

10. To which school would you assign a recruit if you think his chances of success are

	Sonarman		Electronics Technician
(a)	70% _____	and	60% _____
(b)	70% _____	and	70% _____
(c)	70% _____	and	80% _____
(d)	70% _____	and	90% _____
(e)	70% _____	and	95% _____

11. To which school would you assign a recruit if you think his chances of success are

	Hospital Corpsman		Storekeeper
(a)	80% _____	and	60% _____
(b)	80% _____	and	70% _____
(c)	80% _____	and	80% _____
(d)	80% _____	and	90% _____
(e)	80% _____	and	95% _____

12. To which school would you assign a recruit if you think his chances of success are

	Engineman		Disbursing Clerk
(a)	80% _____	and	60% _____
(b)	80% _____	and	70% _____
(c)	80% _____	and	80% _____
(d)	80% _____	and	90% _____
(e)	80% _____	and	95% _____

13. To which school would you assign a recruit if you think his chances of success are

	Storekeeper		Postal Clerk
(a)	90% _____	and	60% _____
(b)	90% _____	and	70% _____
(c)	90% _____	and	80% _____
(d)	90% _____	and	90% _____
(e)	90% _____	and	95% _____

14. To which school would you assign a recruit if you think his chances of success are

	Disbursing Clerk		Electronics Technician
(a)	80% _____	and	60% _____
(b)	80% _____	and	70% _____
(c)	80% _____	and	80% _____
(d)	80% _____	and	90% _____
(e)	80% _____	and	95% _____

15. To which school would you assign a recruit if you think his chances of success are

	Disbursing Clerk		Yeoman
(a)	80% _____	and	60% _____
(b)	80% _____	and	70% _____
(c)	80% _____	and	80% _____
(d)	80% _____	and	90% _____
(e)	80% _____	and	95% _____

16. To which school would you assign a recruit if you think his chances of success are

Electronics Technician		and	Hospital Corpsman	
(a) 70%	______	and	60%	______
(b) 70%	______	and	70%	______
(c) 70%	______	and	80%	______
(d) 70%	______	and	90%	______
(e) 70%	______	and	95%	______

17. To which school would you assign a recruit if you think his chances of success are

Machinist's Mate		and	Storekeeper	
(a) 90%	______	and	60%	______
(b) 90%	______	and	70%	______
(c) 90%	______	and	80%	______
(d) 90%	______	and	90%	______
(e) 90%	______	and	95%	______

18. To which school would you assign a recruit if you think his chances of success are

Hospital Corpsman		and	Sonarman	
(a) 80%	______	and	60%	______
(b) 80%	______	and	70%	______
(c) 80%	______	and	80%	______
(d) 80%	______	and	90%	______
(e) 80%	______	and	95%	______

19. To which school would you assign a recruit if you think his chances of success are

Yeoman		and	Storekeeper	
(a) 80%	______	and	60%	______
(b) 80%	______	and	70%	______
(c) 80%	______	and	80%	______
(d) 80%	______	and	90%	______
(e) 80%	______	and	95%	______

20. To which school would you assign a recruit if you think his chances of success are

	Sonarman		Radioman
(a)	70% ______	and	60% ______
(b)	70% ______	and	70% ______
(c)	70% ______	and	80% ______
(d)	70% ______	and	90% ______
(e)	70% ______	and	95% ______

21. To which school would you assign a recruit if you think his chances of success are

	Hospital Corpsman		Yeoman
(a)	80% ______	and	60% ______
(b)	80% ______	and	70% ______
(c)	80% ______	and	80% ______
(d)	80% ______	and	90% ______
(e)	80% ______	and	95% ______

22. To which school would you assign a recruit if you think his chances of success are

	Radioman		Disbursing Clerk
(a)	70% ______	and	60% ______
(b)	70% ______	and	70% ______
(c)	70% ______	and	80% ______
(d)	70% ______	and	90% ______
(e)	70% ______	and	95% ______

23. To which school would you assign a recruit if you think his chances of success are

	Yeoman		Engineman
(a)	80% ______	and	60% ______
(b)	80% ______	and	70% ______
(c)	80% ______	and	80% ______
(d)	80% ______	and	90% ______
(e)	80% ______	and	95% ______

24. To which school would you assign a recruit if you think his chances of success are

	Postal Clerk			Machinist's Mate	
(a)	90%	_____	and	60%	_____
(b)	90%	_____	and	70%	_____
(c)	90%	_____	and	80%	_____
(d)	90%	_____	and	90%	_____
(e)	90%	_____	and	95%	_____

APPENDIX D. THE QUESTIONNAIRE USED IN MEASURING THE UTILITY OF GRADUATES BY THE MAGNITUDE ESTIMATION SCALING METHOD

Name______________________Position____________Years Service_______
    Last   First

Years in present position__________Rank/Rate_______________

Years in present command__________

The Estimation of Manpower Needs

<u>What this is about</u>

As you know, there are severe shortages of personnel in certain ratings. We often refer to these ratings as "critical." Other ratings are less critical since the supply of qualified persons in these ratings is more nearly sufficient to meet the requirements of the Navy. Still other ratings have enough men and are not critical at all.

The ratings of the Navy can be thought of as lying on a scale which runs from non-critical at one end to very critical at the other end. A rating's position on this scale would indicate how badly the Navy needs more men in that rating.

One way to determine how critical a rating is, is to ask experts to place the ratings on a numerical scale. "Expert" is defined as someone who knows a great deal about the personnel needs of the Navy. Since you are involved in the distribution of enlisted personnel, you are an expert in this regard.

<u>This is a research effort</u>

You will be asked below to indicate how critical you believe certain ratings to be. We want your own opinion so do not discuss this with others or consult other estimates of personnel needs. Your answers will be used only for research, and will be held confidential.

<u>Avoid short-term fluctuations</u>

Try to base your judgements on an extended time period so as to avoid short-term fluctuations in the need for, and the supply of, men in certain ratings. Use the calendar year 1963 as the period for your judgements.

The task

At this time we are interested in just 10 ratings:

Radioman	______	Hospitalman	______
Disbursing Clerk	______	Sonarman	______
Electronics Technician	______	Postal Clerk	______
Yeoman	______	Storekeeper	______
Machinist's Mate	______	Engineman	______
________________		__0__ (not critical)	

First, write the name of a rating which you believe was not at all critical in the bottom blank. Do not use any of the 10 ratings listed. This rating should represent the zero point of the scale--a rating in which there was an abundance of men.

Next, put the number 100 in the blank at the right of the rating that you believe was the most critical of the 10 listed. Now assign numbers between zero and 100 to the other ratings to show how critical they were in your judgment. But first, understand that these numbers should be chosen in relation to the zero and the 100 which you have already assigned to ratings. Thus, if you think that one of the ratings was exactly half as critical as the one that you chose as most critical, assign to it the number 50; or if you think it was one-fourth as critical, assign to it the number 25, etc. Write the number chosen for each rating on the list in the blank at the right of that rating. In this way you will be placing the ratings on a 100-point scale.

APPENDIX E. SUCCESS FUNCTION CALCULATIONS

TABLE 16

WORKTABLE IN THE COMPUTATION OF THE SUCCESS FUNCTION QUANTITIES

School	N	M_Y	σ_Y	(S) $\frac{65-M_Y}{\sigma_Y}$	p	r	(T) $\frac{S}{r}$	M_x	σ_x	(U) $M_x+\sigma_x T$	(V) $\frac{\sqrt{(1-r^2)}}{r}$	(W) $\sigma_x V$	x_c	$\frac{x_c-U}{W}$	$P(G\|x_c)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
SO	387	72.84	7.30	-1.07	.85	.61	-1.754	101.64	14.89	75.52	1.298	19.327	110	1.78	.96
ET	446	63.55	9.58	.15	.56	.61	0.238	156.10	22.17	161.38	1.232	27.313	170	.32	.62
RM	519	74.78	8.73	-1.12	.88	.55	-2.036	101.64	14.89	71.32	1.518	22.603	100	1.27	.90
YN	417	78.33	7.04	-1.89	.97	.45	-4.200	98.87	12.50	46.37	1.984	24.800	110	2.57	.99
SK	272	75.54	9.60	-1.10	.87	.57	-1.930	101.64	14.89	72.90	1.442	21.471	105	1.50	.93
HM	403	81.11	6.99	-2.30	.99	.62	-3.718	101.64	14.89	46.28	1.266	18.851	100	2.85	1.00
MM	554	78.13	9.71	-1.35	.90	.68	-1.985	101.78	14.67	72.66	1.078	15.814	105	2.04	.97
DK	177	77.81	8.69	-1.28	.90	.52	-2.462	101.64	14.89	65.26	1.645	24.494	105	1.65	.95
EN	214	71.61	8.19	-.81	.80	.63	-1.286	100.90	13.87	83.06	1.232	17.088	105	1.28	.90

Notes: Y denotes the criterion distribution.
x denotes the test distribution.
S, T, U, V, and W denote entries in columns (5), (8), (11), (12), and (13), respectively.
Column (5) entries are the z-scores of the graduate-fail division points on the criteria.
Column (8) entries are means of the success functions.
Column (11) entries are means of the success functions expressed in units of the test scales.
Column (12) entries are σ's of the success functions.
Column (13) entries are σ's of the success functions expressed in units of the test scales.
Column (15) entries are success function z-scores of the cutoffs.
All entries are population values.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) U.S. Naval Personnel Research Activity San Diego, California 92152	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

THE APPLICATION OF DECISION THEORY AND SCALING METHODS TO SELECTION TEST EVALUATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

Ervin W. Curtis

6. REPORT DATE February 1967	7a. TOTAL NO. OF PAGES 112	7b. NO. OF REFS 15
8a. CONTRACT OR GRANT NO. b. PROJECT NO. IER16000001 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Bulletin STB 67-18 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Chief of Naval Personnel (Pers-A3) Department of the Navy Washington, D.C. 20370

13. ABSTRACT

The correlational approach to selection test evaluation was examined and found to have serious limitations. An approach based on statistical decision theory was developed. Two new methods were presented, one called the utility function method and the other the decision-theoretic method. The former involves the comparison of criterion groups in terms of their utility to the institution using the selection test. The decision-theoretic method is based on statistical decision theory and involves the construction of a payoff matrix corresponding to the contingency table relating the test to the criterion. The cell frequencies are weighted in a utility equation by the payoff values in the corresponding cells of the payoff matrix. This utility equation represents a new test evaluation index which directly expresses the utility of the test to the institution using it. Both of these new methods require the measurement of values peculiar to the institution using the test. The utility function method requires that the performance criterion be translated to a utility function; while the decision-theoretic method requires that a payoff matrix be developed which reflects the gains and losses each cell observation represents to the institution. The three methods (correlational, utility function, and decision-theoretic) were compared with tests used to select students for A-Schools in the U.S. Navy. The three methods led to quite different indications regarding the utility of the selection tests evaluated. The two new methods agreed in terms of the proportion improvement over chance prediction provided by the tests while the correlational method tended to underestimate this proportion. In terms of practical significance the decision-theoretic method lead to much more positive conclusions regarding the tests than did the other two methods.

DD FORM 1473 1 NOV 65 (PAGE 1)

S/N 0101-807-6801

UNCLASSIFIED
Security Classification

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT

DD FORM 1 NOV 65 1473 (BACK)
(PAGE 2)

UNCLASSIFIED
Security Classification